

# **Computer Aided Dysplasia Grading for Barrett's Oesophagus Virtual Slides**

**by**

*Afzan Adam*

**Submitted in accordance with the requirements  
for the degree of Doctor of Philosophy.**



**UNIVERSITY OF LEEDS**

**The University of Leeds  
School of Computing**

**May 2015**



**The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included.**

**The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.**

Some parts of the work presented in this thesis have been published in the following articles:

**Adam, A., Bulpitt, A. and Treanor, D.,** "Texture analysis of virtual slides for grading dysplasia in Barretts oesophagus", *Proceedings of Medical Image Understanding and Analysis*, pp.269-273. London, July 2011.

This experimental results are credited to Afzan Adam and form the first part of Chapter 4 of this thesis. The data analysis and manuscript editing contributed together with Andrew J. Bulpitt while Darren Treanor contributing on the clinical aspect of the paper.

**Adam, A., Bulpitt, A. and Treanor, D.,** "Grading Dysplasia in Barrett's Oesophagus Virtual Pathology Slides with Cluster Co-occurrence Matrices", *In Proc. of Histopathology Image Analysis: Image Computing in Digital Pathology* in conjunction with The 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). Nice, October 2012.

This experimental results are credited to Afzan Adam and form the second part of Chapter 4 of this thesis. The data analysis and manuscript editing contributed together with Andrew J. Bulpitt while Darren Treanor contributing on the clinical aspect of the paper.

**This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.**

**@2015 The University of Leeds and Afzan Adam.**

# Acknowledgements

*Alhamdulillah*, I have managed to complete my thesis with support from many people around me. I would never be able to mention everyone names here but the motivation, support and helps received will never be forgotten.

My utmost gratitude are for my research supervisor; Dr. Andrew J. Bulpitt who has directly guide, motivate and educate me with research skills. Also to both of my examiners, for their critical feedbacks to improve my thesis quality. My gratitude also goes to:

- Dr Darren Treanor, a Consultant Pathologist at St. James's Hospital for his expertise, guidance and teachings.
- My financial sponsor; Ministry of Higher Education, Malaysia and Universiti Kebangsaan Malaysia.

Without them, I will not be able to continue my study and complete this research.

Also to the expert pathologist, Dr Nigel Scott and Prof Mike Dixon at Leeds Teaching Hospitals Trust. As well as the Machine Vision Group at School of Computing, Leeds, especially Virtual Pathology members, Dr DR Magee, Joe, Alex, Yi and Yu.

Not forgetting the dearest ones to me who has sacrifices a lot including their career, time, money and/or emotions:

- my husband, my son, my daughter and my baby: you are the sources of my strength
- my parents and siblings: who has never doubted my ability and it kept me going
- my dearest friends in lab, in Leeds and back in Malaysia: who kept me insane

And, definitely not the least are Sammy & Noushin, Kak Nita and Leo, all the support people for ARC1 (Dureid El-Moghraby), virtualpathology server at LIMM (Dave Turner & Mike Hale) , VPN Leeds , IT Centre UKM(Tasneem) and everyone who has touch my life as a postgraduate student at Leeds University, UK. ...

*Thank  
You*

# Abstract

Dysplasia grading in Barrett's Oesophagus has been an issue among pathologist world-wide. Despite of the increasing number of sufferers every year especially for westerners, dysplasia in Barrett's Oesophagus can only be graded by a trained pathologist with visual examination.

Therefore, we present our work on extracting textural and spatial features from the tissue regions. Our first approach is to extract only the epithelial layer of the tissue, based on the grading rules by pathologists. This is carried out by extracting sub images of a certain window size along the tissue epithelial layer. The textural features of these sub images were used to grade regions into dysplasia or not-dysplasia and we have achieved 82.5% AP with 0.82 precision and 0.86 recall value. Therefore, we have managed to overcome the 'boundary-effect' issues that have usually been avoided by selecting or cropping tissue image without the boundary.

Secondly, the textural and spatial features of the whole tissue in the region were investigated. Experiments were carried out using Grey Level Co-occurrence Matrices at the pixel-level with a brute-force approach experiment, to cluster patches based on its texture similarities. Then, we have developed a texture-mapping technique that translates the spatial arrangement of tissue texture within a tissue region on the patch-level. As a result, three binary decision tree models were developed from the texture-mapping image, to grade each annotated regions into dysplasia Grade 1, Grade 3 and Grade 5 with 87.5%, 75.0% and 81.3% accuracy percentage with kappa score 0.75, 0.5 and 0.63 respectively.

A binary decision tree was then used on the spatial arrangement of the tissue texture types with respect to the epithelial layer to help grade the regions. 75.0%, 68.8% and 68.8% accuracy percentage with kappa value of 0.5, 0.37 and 0.37 were achieved respectively for dysplasia Grade 1, Grade 3 and Grade 5. Based on the result achieved, we can conclude that the spatial information of tissue texture types with regards to the epithelial layer, is not as strong as is on the whole region.

The binary decision tree grading models were applied on the broader tissue area; the whole virtual pathology slides itself. The consensus grading for each tissue is calculated with positivity table and scoring method. Finally, we present our own thresholded frequency method to grade virtual slides based on frequency of grading occurrence; and the result were compared to the pathologist's grading. High agreement score with 0.80 KV was achieved and this is a massive improvement compared to a simple frequency scoring, which is only 0.47 KV.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Glass and virtual pathology slides . . . . .	1
1.2	Barrett’s Oesophagus . . . . .	2
1.3	Research motivation, objectives and contributions . . . . .	4
1.4	Research framework and thesis outline . . . . .	6
<b>2</b>	<b>Literature Review</b>	<b>9</b>
2.1	Clinical challenges . . . . .	10
2.2	Virtual pathology slides . . . . .	13
2.3	Image pre-processing . . . . .	14
2.4	Tissue texture analysis . . . . .	15
2.4.1	Statistical feature extraction techniques . . . . .	16
2.4.2	Spectral feature extraction techniques . . . . .	24
2.5	Tissue architecture analysis . . . . .	26
2.6	Machine learning to enable pathologist-like diagnosis. . . . .	31
2.6.1	Supervised learning . . . . .	31
2.6.2	Unsupervised learning . . . . .	35
2.6.3	Semi-supervised learning . . . . .	36
2.7	Conclusions . . . . .	37
<b>3</b>	<b>Tissue detection and selection process</b>	<b>38</b>
3.1	Ground truth: Whole virtual slides . . . . .	39
3.2	Ground truth: Annotated regions . . . . .	41
3.3	Noise reduction . . . . .	42
3.4	Region creation . . . . .	44
3.4.1	Curvature-based regions . . . . .	44
3.4.2	Tile-based regions . . . . .	47
3.5	Colour Normalisation . . . . .	48

3.5.1	Normalisation method . . . . .	49
3.5.2	Result and conclusions . . . . .	51
3.6	Conclusions . . . . .	53
<b>4</b>	<b>Annotated Region Tissue Analysis</b>	<b>54</b>
4.1	Epithelial layer texture analysis . . . . .	55
4.1.1	Reference point selection . . . . .	56
4.1.2	Patch creation . . . . .	59
4.1.3	Feature extraction . . . . .	62
4.1.4	Results and discussion . . . . .	65
4.2	Tissue texture analysis . . . . .	66
4.2.1	Cluster co-occurrence images . . . . .	69
4.2.2	The Cluster Co-occurrence Matrices . . . . .	70
4.2.3	Result . . . . .	73
4.2.4	Discussion . . . . .	77
4.3	Spatial feature analysis . . . . .	78
4.3.1	Results and conclusions . . . . .	78
4.4	Conclusions . . . . .	80
<b>5</b>	<b>Diagnosing whole virtual slides</b>	<b>82</b>
5.1	Virtual slides preparation . . . . .	83
5.2	Tissue and region selection . . . . .	84
5.3	Implementation of epithelial layer analysis . . . . .	85
5.4	Implementation of BDT models . . . . .	86
5.5	Virtual slides grading . . . . .	87
5.6	Conclusions . . . . .	91
<b>6</b>	<b>Conclusion and future work</b>	<b>92</b>
6.1	Summary of work . . . . .	92
6.2	Discussion and Future Work . . . . .	94
<b>A</b>	<b>Colour normalisation result</b>	<b>96</b>
<b>B</b>	<b>Epithelial layer analysis</b>	<b>100</b>
<b>C</b>	<b>Texture and spatial mapping: CCI</b>	<b>103</b>



# List of Figures

1.1	Region of BO in human body (a), sample of normal oesophagus tissue lining (b) and sample of oesophagus lining with Low Grade Dysplasia of BO (c) . . . . .	3
1.2	Framework of analysing and diagnosing dysplasia in BO . . . . .	8
2.1	Images of tissue from a virtual pathology slide in different zoom level. The green boxes show the area which is being zoomed. . . . .	10
2.2	Sample of images with different texture. Note the entropy values for different images (a-c), but same image with different scaling produce higher difference (b,d,e). . . . .	16
2.3	Matrices operation to produce symmetrical matrix. . . . .	17
2.4	Asymmetrical GLCM generated from sample image. . . . .	18
2.5	MCM channel pairwise in RGB images. . . . .	18
2.6	The different texture of epithelial surface compared to lamina region in oesophagus tissue. The four glcm features values are shown in the graphs provided. . . . .	20
2.7	Matrix operation to calculate LBP value . . . . .	23
2.8	Sample of tissue architecture model . . . . .	29
2.9	Sample of oesophagus tissue with tissue surfaces, basement membrane, epithelial layer and the lamina propria. (left) Sample of normal tissue with wider epithelial while (right) is sample of tissue with dysplasia grade 3. Narrower lamina layer and curvier profile of the basement membrane can be seen. . . . .	30
2.10	Optimum hyperlane projected by SVM learning algorithm for linear classification prblem. The points denoted with (*) is the support vectors, which is the closest points to the hyperlane. . . . .	34

2.11	(a), (b), and (c), shows the same two class separation problem, but in three different view. Linear separation in for data points in input space is not possible with minimum erros (a) but a better separation is possible with a non-linear surface (b). This non-linear surface is in-fact a linear separation line when viewed from the feature space (c). (Image is curtosy from [81] ) . . . . .	35
3.1	Slide 11013 with consensus grading as grade 4 but regions annotated marked as grade 4, grade 2 and grade 4 respectively. . . . .	38
3.2	Relationship of virtual slides and annotated region training and test set. . . . .	41
3.3	Graph plot showing number of detected objects with different values of eliminated area ( $\tau$ ). . . . .	43
3.4	Image de-noising process . . . . .	43
3.5	Image showing the tissue shape representative with different window size $j$ . $j=20$ is selected and the pixel along the image bounding box is eliminated. . . . .	45
3.6	Implementation of region creation. . . . .	46
3.7	Tissue classification based on <i>hcp</i> thresholding. The first two images from left are classified as smooth, and the last two are complex. . . . .	47
3.8	Tissue regions created based on complex and smooth boundaries from <i>hcp</i> . Note the overlapping areas of region 1 and 2 in region 3. . . . .	47
3.9	Non-tissue artefacts commonly found in virtual pathology slides. . . . .	48
3.10	Second method of region creation, with selected region. . . . .	49
4.1	Region for texture extraction from annotated regions . . . . .	54
4.2	Two different baselines to choose as a reference point in creating patches. . . . .	56
4.3	Sample of mis-detected boundary and the corrected output. . . . .	57
4.4	Colour deconvolution process to detect nuclei boundary as a reference point to create patches. . . . .	58
4.5	Sample annotated regions with three different boundaries. . . . .	59
4.6	(a) shows the creation of rotated patches along the detected epithelial layer while (b) shows the un-rotated patches. $R$ represent the height and $\theta$ represents the angle for rotation. . . . .	60
4.7	Figure showing the rotated and un-rotated patches with different $R$ , created with the algorithm shown. . . . .	60
4.8	Grading annotated regions into G1 and G5 using different set of GLCM features, $k$ and $pz$ . . . . .	63



4.9	Graphs showing the AP and KV achieved with different sets of features, to test the effect of rotation for patches. . . . .	63
4.10	Texture features extraction steps from the lamina propria. . . . .	67
4.11	Sample of clustered patches with $pz=100*100$ , zoom=20X and $n=10$ from EXP1. . . . .	68
4.12	CCI generated from EXP1 and EXP2. We can see the difference between images produced from both experiments. . . . .	70
4.13	Comparison of binary tree grading result between different sets of textures in EXP1 and EXP2. . . . .	74
5.1	Processs involved in grading a virtual slide. . . . .	83
C.1	Sample CCI with different $k$ values. The zoom level is set to 20X, $pz=100*100$ pix and $n=4$ . . . . .	104
D.1	BDT model selected for grading CCI into G1 or not-G1, using feature A7 from EXP2. . . . .	107
D.2	BDT model selected for grading CCI into G3 or not-G3, using feature set A3 from EXP2. . . . .	108
D.3	BDT model selected for grading CCI into G5 or not-G5, using feature A1 from EXP1. . . . .	109
D.4	Spatial BDT model for G1 vs not-G1. . . . .	110
D.5	Spatial BDT model for G3 vs not-G3. . . . .	111
D.6	Spatial BDT model for G5 vs not-G5 is not balanced and the root node starts with C5. . . . .	112

# List of Tables

1.1	Comparison of classification for nomenclature of the Barrett's associated dysplasia. Adapted from [87]	4
2.1	Showing sample agreement between two observers. $a$ and $d$ is total number of agreement between observer 1 and observer 2, while $b$ and $c$ is the number of disagreement. $N$ is total number of cases being observed.	11
2.2	Cytological and architectural changes in BO disease across stages of dysplasia. Note that -:absent; +/-:maybe present; +: usually present	12
3.1	Range of dysplasia grading, which could be used.	40
3.2	Available virtual slides, training and testing set according to grades.	41
3.3	Excluded annotated regions.	42
3.4	Parameter setting for complexity measurement.	46
3.5	Images used to test combinations of different normalisation techniques with different colour classifier models.	49
3.6	Normalisation failure in additional images	52
3.7	Bimodal normalisation failures on ground truth images	53
4.1	Validation result on values of $n, k$ and zoom level to grade annotated regions into G1 or G6 using texture of tissue boundary.	64
4.2	Average result with 10-fold cross validation for comparison of rotated and un-rotated patches with $p_z = 150$ .	64
4.3	Comparison of grading annotated regions into dysplasia and non-dysplasia between tissue texture along nuclei lining and along tissue boundary with $k = 5$ .	65
4.4	Parameters tested and selected values for epithelial layer analysis.	66
4.5	Grading CCI with decision tree into G1, G3 and G5. Current agreement between pathologist is 0.24KV.	68
4.6	Significant clusters of textures to classify certain grade.	69

4.7	CCM textural features derived from CCI. . . . .	71
4.8	AP for testing results on values of $n$ , $k$ and zoom level to grade annotated regions into G1 or G6 using CCM features. . . . .	71
4.9	Comparison of classification between Decision Tree and Random Forest achieved at 20X. (* not found in literature) . . . . .	72
4.10	Grading AP with $pz=100*100$ on different values of $k$ . . . . .	73
4.11	Tested and selected value for parameter setting for whole annotated region texture analysis . . . . .	73
4.12	Test result for CCM features selection from EXP1 and EXP2. . . . .	74
4.13	Grading performance of the selected CCM features to grade dysplasia in regions into G1, G3 or G5. . . . .	76
4.14	The grading performance of CCM features on a validation data with SVM, RF and BDT . . . . .	77
4.15	The gading performance of spatial features with SVM, RF and BDT . . . .	79
5.1	List of virtual slides test data with the consensus achieved for glass and virtual slides diagnosis. . . . .	84
5.2	Table show sample of tissues detected and filtered, as well as the regions accepted for region grading process from a virtual slide using our selected parameters and threshold values. . . . .	85
5.3	Positivity table and the score used to find a consensus diagnosis for a region. . . . .	87
5.4	Grading score gained for regions in tissue number 2 of slide 13348. . . . .	88
5.5	Proportion of regions detected with G3 and G5 for each class of virtual slides. . . . .	89
5.6	Confusion matrix between our grading method and the ground truth (glass slide diagnosis). . . . .	89
5.7	Grading result for virtual slides using our CCM features and thresholded frequency value. $f_{G1}$ is the frequency of G1; $f_{G3}$ is the frequency of G3; $f_{G5}$ is the frequency of G5; $f_tG3$ and $f_tG5$ is the threshold frequency of each grade accordingly; D is our suggested grading with threshold frequency; E is the consensus grading by two pathologists on the glass slides; F is the consensus grading by two pathologists on virtual slides and G is suggested grading based on frequency only. . . . .	90
A.1	Normalisation output from combination of different normalisation techniques and colour classifiers. . . . .	97
A.2	Normalised images . . . . .	98

A.3	Normalised images continued . . . . .	99
B.1	sample AR with clustered rotated patches on detected tissue boundary. . .	101
B.2	Sample AR with clustered unrotated patches on detected tissue boundary.	102
C.1	CCI from EXP1 and EXP2 with different magnification. . . . .	105

# Abbreviations

<b>H&amp;E</b>	<b>H</b> ematoxylin and <b>E</b> osin
<b>BO</b>	<b>B</b> arrett's <b>O</b> esophagus
<b>LGD</b>	<b>L</b> ow <b>G</b> rade <b>D</b> ysplasia
<b>HGD</b>	<b>H</b> igh <b>G</b> rade <b>D</b> ysplasia
<b>IMC</b>	<b>I</b> ntra <b>M</b> ucousal <b>C</b> arcinoma
<b>KV</b>	<b>K</b> appa <b>V</b> alue
<b>AP</b>	<b>A</b> ccuracy <b>P</b> ercentage
<b>DAB</b>	<b>D</b> iaminobenzidine
<b>GLCM</b>	<b>G</b> rey <b>L</b> evel <b>C</b> o-occurency <b>M</b> atrix
<b>MCM</b>	<b>M</b> ultiple <b>C</b> hannel <b>M</b> atrix
<b>RGB</b>	<b>R</b> ed <b>G</b> reen <b>B</b> lue
<b>HSV</b>	<b>H</b> ue <b>S</b> aturation <b>V</b> alue
<b>OD</b>	<b>O</b> ptical <b>D</b> ensity
<b>KNN</b>	<b>K</b> Nearest <b>N</b> eighbour
<b>SVM</b>	<b>S</b> upport <b>V</b> ector <b>M</b> achine
<b>CT</b>	<b>C</b> omputed <b>T</b> omography
<b>BoW</b>	<b>B</b> ag of <b>W</b> ords
<b>GI</b>	<b>G</b> astro <b>I</b> ntestinal
<b>LBP</b>	<b>L</b> ocal <b>B</b> inary <b>P</b> attern
<b>PCA</b>	<b>P</b> rincipal <b>C</b> omponent <b>A</b> nalysis
<b>SIFT</b>	<b>S</b> cale <b>I</b> nvariant <b>F</b> eature <b>T</b> ransform
<b>HoG</b>	<b>H</b> istogram of <b>O</b> rientation <b>G</b> radient
<b>MRI</b>	<b>M</b> agnetic <b>R</b> esonance <b>I</b> maging
<b>VT</b>	<b>V</b> oronoi <b>T</b> esellations
<b>MST</b>	<b>M</b> inimum <b>S</b> panning <b>T</b> ree
<b>CINN</b>	<b>C</b> ervical <b>I</b> Ntraepithelial <b>N</b> eoplasia
<b>GC</b>	<b>G</b> aussian <b>C</b> lassifier
<b>DT</b>	<b>D</b> ecision <b>T</b> ree
<b>RF</b>	<b>R</b> andom <b>F</b> orest
<b>hcp</b>	<b>h</b> igh <b>c</b> urvature <b>p</b> oints
<b>CCI</b>	<b>C</b> luster <b>C</b> o-occurence <b>I</b> mages
<b>CCM</b>	<b>C</b> luster <b>C</b> o-occurency <b>M</b> atrix
<b>BDT</b>	<b>B</b> inary <b>D</b> ecision <b>T</b> ree
<b>NC</b>	<b>N</b> o <b>C</b> onsensus

# Chapter 1

## Introduction

---

In the last few decades, we can see that computer applications have been seamlessly combined with many other fields, particularly in education, architecture, design, entertainment and medicine. In medicine, computer applications are used to detect any signs of disease at the earliest stage possible so as to save lives. While medical experts are looking for preventive steps, engineers and computer scientists are trying very hard to automate certain processes in order to speed up, standardize or upgrade the diagnostic process. These offer cheaper, faster and more quantifiable analysis of diseases.

This chapter contains the clinical and technical explanation of the glass and virtual pathology slides as well as the disease that we are looking at; the Barrett's Oesophagus. In continuation to that, the gap in the current practise in diagnosing this disease will be highlighted, as it became our research motivations, followed by our research objectives and contributions. Then, the research framework and thesis outlined will be presented.

### 1.1 Glass and virtual pathology slides

One of the standard procedures to diagnose certain conditions is histopathological examination; diagnosis based on the visual observation for changes of patterns, shapes and sizes of tissue structures from tissue samples called biopsies. These tissue samples were preserved in pathology glass slides. The pathology glass slides are prepared by a series of steps, starting by taking biopsy samples from the patient. The biopsies are then placed in

a sample cassette and put into a special device to replace all water inside the biopsies with paraffin. Afterwards, the sample will be solid and surrounded in a cube of wax where it will be cut with a microtome into very thin slices ( $\pm 5\mu\text{m}$ ). These slices are then placed on a glass slide for Hematoxylin and Eosin (**H&E**) or other staining. Lastly, another thin glass known as a cover slip is glued on top of the biopsy glass.

Pathology slides are then sent to pathologists who are trained to notice any abnormalities and make a visual evaluation of the tissue samples. To help visualize the cell structures and identify any abnormalities in the tissue, many types of dye or staining are used for contrasting colours. However, the tissues slices vary in number, size, thickness, shape and also staining concentrations. Glass slide diagnosis is used as a gold standard procedure, but it has many drawbacks as slides are fragile, not easily transportable for presentation, and the staining may fade or even disappear over time.

Therefore, digital scanner technology which is currently capable of producing high resolution images in a short time has become a springboard for the virtual pathology slide diagnosis evolution. Virtual pathology slides are glass slides which have been scanned and digitized with specialized digital scanner technology. It offers many advantages as they are easily saved, archived and retrieved without compromising the image quality. The slides are more interactive, easily shared, presented, standardized and annotated as well as presented for education, discussion and visual evaluation purposes. It enables the development of automatic or semi-automatic detection where changes or abnormalities in tissue structure can be detected, measured and analysed in a quantitative ways. Virtual pathology slides have more advantages to revolutionize the way diagnosis is carried out in pathology, replacing the microscope based method.

## 1.2 Barrett's Oesophagus

Barrett's Oesophagus (**BO**), (sometimes referred as Barrett's ulcer or columnar lined oesophagus) is a pre-malignant but treatable condition [90] where '*any portion of the normal squamous lining has been replaced by a metaplastic columnar epithelium that is visible microscopically*' [44]. However, the American College of Gastroenterology, German Society of Pathology, Amsterdam Working Group and French Society of Digestive Disease has included the histological evidence of intestinal metaplasia in their definition of BO [75]. Intestinal metaplasia is the existence of abnormal characteristics of both stomach tissue and intestinal tissue in the oesophagus, as an adaptive response of the normal oesophagus cells towards the acid reflux coming from the stomach. Rather than transforming into squamous cells, they mature into mucin-producing columnar cells, which are

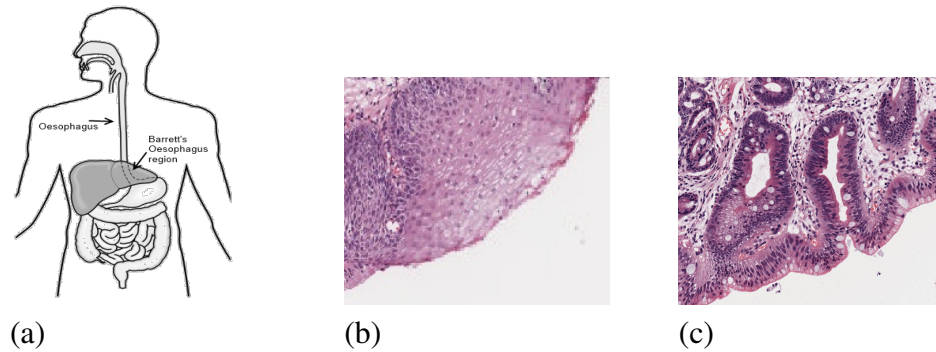


Figure 1.1: Region of BO in human body (a), sample of normal oesophagus tissue lining (b) and sample of oesophagus lining with Low Grade Dysplasia of BO (c)

better able to protect themselves from the acidic environment [5].

The series of changes from normal squamous cells with a smooth surface as in Figure 1.1(b), to metaplastic columnar cells as in Figure 1.1(c), usually starts with the formation of mucus secreting goblets and glands in the epithelium layer of oesophagus tissue. Then, the same formation appears in the lamina propria layer of the tissue, and noticeable changes in size, shape and other cytological features occur in the cells and nuclei. These changes are generally labelled according to severity; starting with BO, indefinite for dysplasia, low grade dysplasia (**LGD**), high grade dysplasia (**HGD**) and intramucosal carcinoma (**IMC**).

Dysplasia, or intra-mucosal neoplasia is the earliest (pre-invasive) form of cell abnormality and not necessarily cancerous. It is defined as '*neoplastic epithelium confined by the underlying basement membrane of the gland from which it arises*'[99]. Currently in BO, dysplasia can only be graded by a trained pathologist with visual examination of glass slides under a microscope. The level of BO severity is in a continuous pattern, thus differentiating between each type with its characteristics is not an easy task. The interobserver agreement between pathologists for grades of dysplasia in BO is only 57% [119]. Two classifications used for grading dysplasia are recommended by Dysplasia Morphology Study Group Classification (routinely used in the United States) and Vienna Classification (widely used in Europe and Asia) is shown in Table 1.1.

Despite of the conflicting definitions and diagnostic challenge, grading BO is very important as it is one of the earliest key sign for cancer possibility. BO sufferers are estimated to have increased to about 1-5% of the western population [1]. Furthermore, endoscopic surveillance is mainly performed to identify and monitor the progression of dysplasia in BO and to monitor the progression. However, the diagnosis is becoming more difficult when dysplasia is becoming more severe [131]. On top of that, the risk



Dysplasia Morphology Study group Classification	Vienna Classification
Negative	Negative for neoplasia/dysplasia
Indefinite for Dysplasia	Indefinite for neoplasia/Dysplasia
Low-grade Dysplasia	Non-invasive low grade neoplasia (low grade adenoma/dysplasia)
High-grade Dysplasia	Non-invasive high grade neoplasia
	High-grade dysplasia/adenoma
	Non invasive carcinoma (carcinoma in situ)
	Suspicious for invasive carcinoma
Adenocarcinoma	Invasive neoplasia
Intramucosal	Intramucosal adenocarcinoma
Invasive	Submucosa carcinoma or beyond

Table 1.1: Comparison of classification for nomenclature of the Barrett’s associated dysplasia. Adapted from [87]

of Oesophageal Cancer for the United Kingdom population has increased threefold since 1971 [63] and only 5% of BO’s patients seek medical attention for symptoms of cancer [24].

Although not many patients with BO developed cancer [16], Spechler in [111] has concluded that progression from LGD to HGD/IMC within 5 years ranges between 10-28%, and the risk for HGD to further develop into Oesophagus Adenocarcinoma is approximately 30-50 times more; as reported in [110, 121] or 30-125 times more as in [9]. On the other hand, Solaymani [110] reported that LGD are 10 times more likely to develop into Oesophageal Cancer than Oesophagus Adenocarcinoma, or 0.5% as reported in [106].

### 1.3 Research motivation, objectives and contributions

This research is driven mainly by the interest in bridging the gap in the variations of grading dysplasia in BO. The existence of virtual pathology slides and interest in automatic diagnosis from histopathology and pathology images has fuelled the motivation toward this research.

The interobserver agreement between pathologists in diagnosing dysplasia in BO has always been an issue with the overall agreement ranging from fair to good. The lack of a universally accepted definition of BO amongst the pathology and gastroenterology societies around the world [75] has certainly contributed to the low agreement score. In

addition, the pathologist's own experience in diagnosing dysplasia will certainly influence their grading decisions.

Therefore, we need to help find the pattern or texture difference between tissues in each grade of dysplasia. Dysplasia criteria are varied and the changes of cytological and morphological changes are continuous and smooth, making differentiating between each type is a difficult task. Even experienced gastrointestinal pathologists frequently disagree on the diagnosis of HGD and IMC [91].

Hence, the objective of this research is to help identify and measure dysplasia in BO virtual pathology slides using textural features and spatial relationships. Textural features of BO tissues were extracted from tissue images in virtual pathology slides and the suitable ones were selected. The morphological changes were identified and used to discriminate grades of dysplasia. Using machine vision, machine learning and virtual pathology slides, these changes in tissue might form a model or map for reference. Therefore, the agreement and confidence level in diagnostic results will increase.

In this research, our main contribution is to grade the severity of dysplasia condition in Barrett's Oesophagus into Grade 1, Grade 3 or Grade 5, using image processing and machine learning technique.

In order to do this, several research contributions has been made:

- The solution for 'border effect' suffered by BO, colon, breast, prostate and oral tissue (among others) during digital image processing. This is carried out by detecting the tissue surface, and extracting texture features along the detected surface only.
- The texture-mapping technique named as the Cluster-coded Co-occurrence Image(CCI). This techniques maps the co-existence of many types of tissue textures within a region without compromising its textural information at pixel-level.
- The understanding of spatial arrangement of tissue texture types with reference to the epithelial layer. This knowledge is significantly important as it translated the pathologist's knowledge in examining tissue condition for grading purposes.
- The grading models applied on the whole BO Virtual Slides. This is another contribution to the machine learning and pathology society as the closest research stops at locating dysplastic areas in colorectal tissues.
- Finally, the method to achieve a consensus grading for a virtual slides using grades from each extracted regions. This is achieved by implementing a positivity table that count in the support value gained from grades of each regions extracted.

The outcome of this research has the potential to provide more information for pathologists to challenge or support their grading decisions. They can also use the knowledge from this research to help in teaching, diagnosing and visualizing the reasoning behind each grading with quantifiable features.

The computer vision, image processing and machine learning community can also benefited from this research as the CCI methods might also be used in other domains as well. This research enable us to identifying patterns and rules from tissue images, which might be too subtle for human perception.

## **1.4 Research framework and thesis outline**

Research and clinical papers that have been published related to tissue analysis, especially in BO and other digestive organs such as the mouth, oesophagus and colon, have been referred to to ensure that our research is useful, in demand and novel. It has also been a good springboard for this research for collecting the useful information and tips on the features and methods that have been investigated or attempted before. These are reviewed in Chapter 2 where key papers are discussed in relation to the current work.

Based on the literature review and experiments that were carried out, the developmental framework for a diagnosis support tool has been developed as illustrated in Figure 1.2. The framework shows that the pattern analysis starts with tissue texture and spatial features extraction, analysis, feature selection and decision making model from annotated regions. The implementation is carried out in much larger scale which is on the whole virtual pathology slide.

Chapter 3 contains the detailed explanations of the ground truth data and images available to use. Image pre-processing to reduce noise and normalised the colour in the image are also discussed in this chapter. The main challenge here is to measure the tissue surface complexity, in attempts to divide tissue into meaningful regions. The work has been presented in the Symposium of Microscopy Image Analysis for Biomedical Applications at London, in April 2010. The final outcome from this chapter is the best method to create regions of tissue from the whole tissue extracted from the virtual pathology slides.

Chapter 4 contains the works involved in extracting the texture and spatial features from the annotated region. Annotated regions are smaller regions of the whole tissue image, which are already annotated by pathologists. The texture analysis is divided into three main focuses; 1) the epithelial layer only, 2) the tissue texture features 3) the spatial features between tissue texture arrangement with regards to the epithelial layer. The main challenge in this chapter is the brute-force approach in selecting best feature combina-

tions, and also to come out with a grading model to grade dysplastic regions. The best feature sets are used to grade dysplasia and the result is evaluated, reported, and published in [2] and [3].

Chapter 5 explains the implementation of region creation from Chapter 3, as well as the grading models learnt from Chapter 4 on the whole tissue image extracted from virtual pathology slides. As the virtual pathology slides contain much tissue and each tissue will be divided into regions for dysplasia grading, the main contribution from this chapter is to combine the grading models, as well as to come up with a consensus diagnosis for virtual pathology slide. The consensus diagnosis is evaluated against the pathologist's consensus grading.

Finally, Chapter 6 contains a wrap-up discussions and the conclusions about the research outcomes. It also contains suggestions for possible improvement steps, as well as the possibility to extend the research into commercial use in clinics or education.

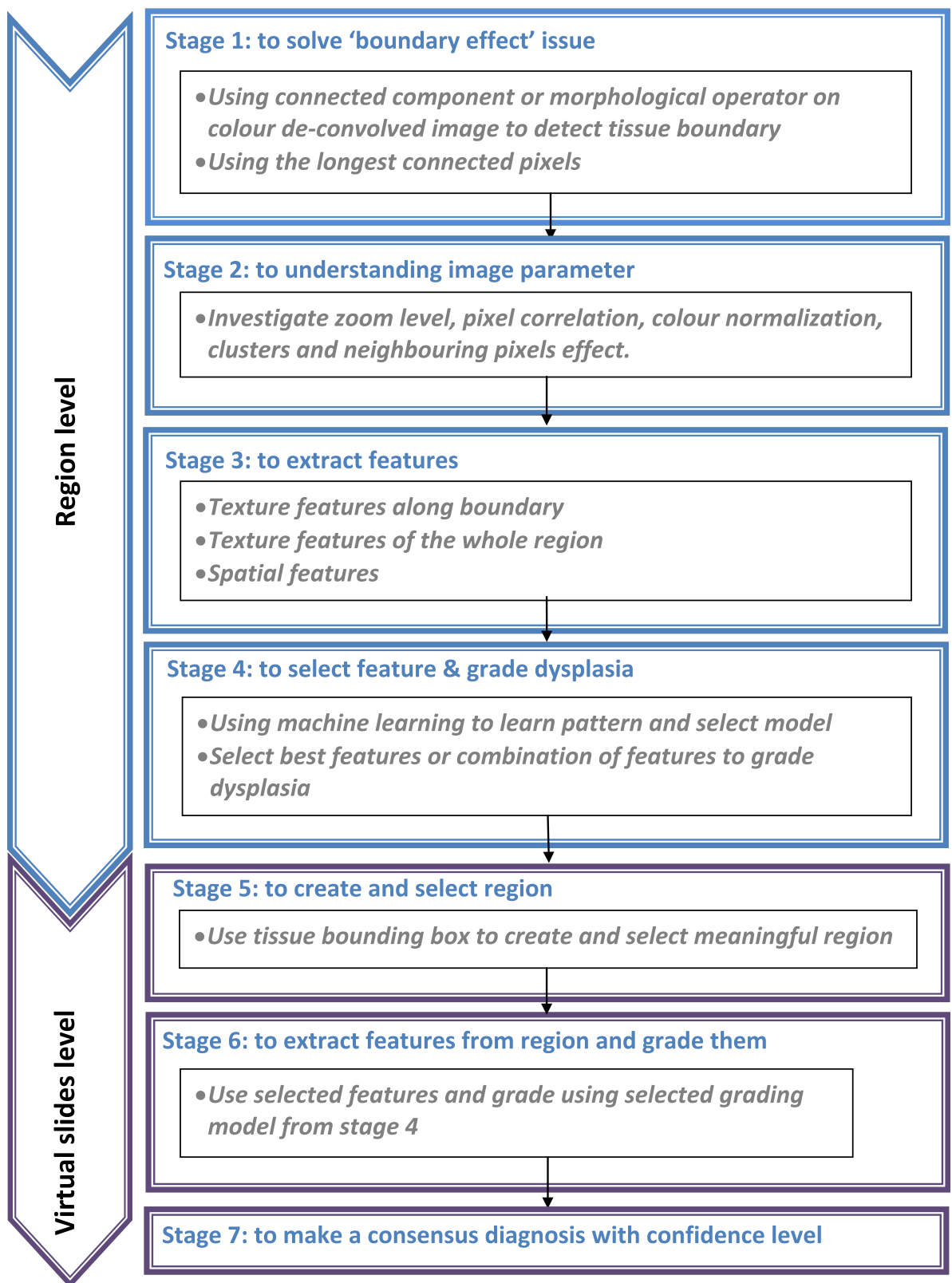


Figure 1.2: Framework of analysing and diagnosing dysplasia in BO

## Chapter 2

# Literature Review

---

Although virtual pathology slides can be of few gigapixels, the growing technology in powerful computing processors, high capacity of hard disks and virtual memory has enabled fast processing regardless of the image size. Thus, virtual pathology slides diagnosis can be carried out at a pathologist's personal desktop computer, where it can be panned and zoomed using a keyboard and a mouse. It also enables simple visualisation tracking, annotation recording for evaluation and other interactivity during diagnosis which is not possible by manual diagnosis using a microscope. The use of suitable computer vision technologies and machine learning techniques have a potential to develop into a computer aided diagnosis system.

Therefore, this chapter will describe the technical aspect of the clinical challenges in grading dysplasia in BO, and how virtual pathology slides has contributed and used in diagnosing and teaching. Then we will look at image pre-processing on H&E images, before continuing on to the texture and architecture features in tissue analysis, as well as attempts by previous researchers to grade or diagnose diseases using machine learning techniques. At the end of this chapter, we will conclude with theories and techniques to be further investigated for this research.

## 2.1 Clinical challenges

Oesophagus tissue samples are obtained from biopsy during endoscopy. In the process, sample tissues are pinched off with a biopsy forceps and sent to a lab for slide preparation including staining, before being analysed by the pathologists. Usually for detecting BO, staining with H&E is enough to contrast between features so the pathologists can see clearly the changes in tissue using a light microscope. However, the abnormalities in dysplasia (as explained in chapter 1.1) form in a continuous spectrum, thus the boundary between negative, indefinite, LGD, and HGD cannot be sharply defined [38].

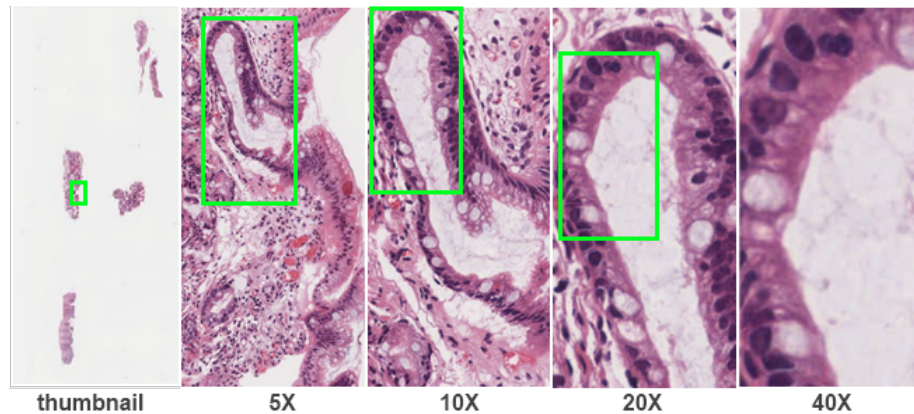


Figure 2.1: Images of tissue from a virtual pathology slide in different zoom level. The green boxes show the area which is being zoomed.

Dysplastic tissue is a combination of many complex structures and pathologists look for abnormalities and changes in tissue to decide on the dysplasia grade. The changes in tissue vary over time and diagnosis becomes more difficult at a higher stage due to its complexity [131]. The degree of the complexity is portrayed by much ongoing clinical research to study the variability of diagnosing dysplasia in BOs, measuring the interobserver and/or intraobserver agreement amongst pathologists, and this has become our research motivation, as explained in chapter 1.2. Some of them can be found in [16, 27, 34, 56, 75, 83, 97].

Interobserver agreement is high when different observers view the same material and have similar interpretations and gradings; while intraobserver agreement counts only when a single observer views the same material on separate occasions and both interpretations are consistent with each other. The agreement is usually measured with kappa value (**KV**), most commonly used in medical literature [120]. KV measure inter-rater agreement for two or more observer for qualitative data. Sample to calculate the agreement between 2 observers is shown in Equation 2.1a, with reference to labels made in

Table 2.1.

	Observer 1			
		yes	no	total
Observer 2	yes	TP	FN	$m_1$
	no	FP	TN	$m_2$
	total	$n_1$	$n_2$	N

Table 2.1: Showing sample agreement between two observers.  $a$  and  $d$  is total number of agreement between observer 1 and observer 2, while  $b$  and  $c$  is the number of disagreement.  $N$  is total number of cases being observed.

$$KV = \frac{P_o - P_e}{1 - P_e} \quad \text{where} \quad (2.1a)$$

$$P_o = TP + TN \quad \text{and} \quad (2.1b)$$

$$P_e = \frac{m_2 \times n_2 + m_1 \times n_1}{N \times N} \quad (2.1c)$$

In comparison to KV, accuracy of a system are commonly measured using accuracy percentage **AP**. AP takes only correctly classified cases compared to number of cases  $((TP + TN)/N)$ . Classes has to be equally balanced to qualify AP as a measurement. Another measurements for system's performance are Precision and Recall. Precision is calculated with  $(TP/(TP + FP))$  while Recall is with  $(TP/(TP + FN))$ .

The morphological changes in tissues include the architectural and cytological changes. Architectural changes concern the overall structure and the arrangement of cells and its components (such as the cytoplasm and nucleus) at a certain area. This can be seen at lower resolution images (for example, image 5X in Figure 2.1); as the overall pattern of a tissue such as cells arrangement, general colour, nuclei/cytoplasm ratio, tissue density and the existing and location of certain local structure.

On the other hand, cytological changes are changes of local structure in a tissue and therefore high resolution images will be used to see each cell in detail. A sample image can be seen at image 40X in Figure 2.1, where the nuclei/cytoplasm ratio of the cell itself as well as its nuclei shape, size and location, can be examined. Table 2.2 shows the criteria of cytological and architectural changes, as suggested by experienced pathologists [27, 38, 82, 87, 111], and how it differs in each stage of dysplasia.



condition	normal	Barrett's	LGD	HGD	IMC	parameter to detect	reference
<b>CYTOLOGICAL CHANGES IN LOCAL FEATURES</b>							
increase nuclei/cytoplasm ratio	-	-	+	++	+++	colour analysis, deconvolution	[103, 130]
loss of polarity	-	-	+	++	++	cell detection	[43, 86, 99, 102, 129]
atypical mitosis	-	+/-	+	++	++	nuclei detection	[43, 86, 102, 129]
full thickness nuclei stratification	-	+/-	+	++	++	area(cyto:nuclei) in cell/texture	[61]
goblet cell	-	++	+	+-	+-	goblet detection, calculate area and location	
hyperchromatin	-	-	-	+	++	colour analysis, deconvolution	[98, 99, 103, 130]
multiple nucleoli	-	-	+/-	+/-	+	nuclei detection	[43, 85, 102, 129]
large irregular nuclei	-	-	-	+/-	++	nuclei segmentation (shape, compactness and roundness)	[43, 85, 99, 102, 129]
irregular nuclei contour and variation of size	-	-	+	++	++		[43, 85, 102, 129]
irregular cell size and shape	-	-	+/-	+	++	cell detection (shape, compactness and roundness)	
necrosis/desmoplasia	-	-	-	-	+/-	tissue compactness/texture	[67]
cell maturity	++	+	+/-	-	-	boundary detection and analysis	
glandulars	-	+	++	+++	+++	gland detection	[23, 82, 85, 86, 128]
<b>ARCHITECTURAL CHANGES IN TISSUE</b>							
viliform change	-	+	++	++	++	boundary detection (curvature, energy and perimeter)	
crypt budding/branching	-	+/-	+	+	++	texture analysis	[23, 32, 60]
crowded (back to back) crypts	-	-	+/-	+	++	texture analysis	[40]
intraluminal papilla/ridges	++	+	+	+/-	-	architectural analysis	
crowded glands	-	-	+/-	+	++	texture analysis, tissue compactness	[99]
crypts breach muscularis mucosa	-	-	-	-	+	crypts detection/ architectural analysis	[23, 32, 60, 99]
existence of infiltration	-	-	-	-	+/-	texture/ architectural analysis	[23, 32, 60, 99]
velvet-like surface	+	+	+/-	-	-	texture/ pattern analysis	[4]

Table 2.2: Cytological and architectural changes in BO disease across stages of dysplasia. Note that -:absent; +/-:maybe present; +: usually present

## 2.2 Virtual pathology slides

Pathologists can intuitively diagnose a disease based on their knowledge and experience of judging the severity of tissue transformation through the pathology slides. However, some abnormalities in tissue might be too subtle for human eyes to perceive, and this can be detected and measured with the help of computer vision. Therefore, the virtual slides images must be of a good quality: good image compression, brightness, focus and field of view [52]. The quality of virtual pathology slide images depend on the glass slides preparation and the scanning setups while acquiring the image.

The practicality of virtual pathology slides has sparked a lot of interest, especially in the application and evaluation in teaching and diagnostic support. Some examples of these cover (but are not limited to) cancers [20] and tumours on major organs [52, 53], dentistry [35], gleason grading in prostate [42], haematology [12] and BO [119]. All these have come to the same conclusion that virtual slides are very useful in teaching and assessment. A further study by Fred R. Dee in [19] has revealed that about 50% of pathology courses worldwide already have or are expected to implement virtual slides in their teaching.

In addition, Lundin et. al concluded that the implementation of a virtual microscopy network spanning a large geographical area can minimize the cost of image compression and is economically feasible, as published in [70]. This research has utilized existing academic networks from Finland, Poland, Netherland, Spain and Sweden during the European Congress of Pathology in 2007. Furthermore, virtual slides have been used by the American Board of Pathology examinations for trainee pathologists [88] as well as in reporting at centralised laboratories in United States of America and Australia [62]. Thus, we can conclude that virtual slide diagnosis is rapidly evolving and is beneficial to the community and pathologists as well.

Virtual pathology slides which are stored in an Aperio server at St James's Hospital, Leeds have been used for this study. These BO pathology slides are a selection of 60 cases from Leeds General Infirmary which represent every grade of dysplasia. The glass slides have been technically inspected by a consultant pathologist (Dr. Darren Treanor) and were scanned with Aperio T3 with a 40 times objective lens. These virtual slides digitise at a resolution of 0.23 micron per pixel, zoomable from thumbnail sized images into 40X magnification and panable; similar to conventional microscopy [45]. It also allows annotations and coordinates recording.

## 2.3 Image pre-processing

The existence of virtual pathology slides has enabled tissue textural or structural features extraction [49] at pixel-based level or region-based level. However, there are three main processes in the computerised tissue analysis framework. The first process is the image preprocessing, secondly is the feature extraction and selection, and the final process is the analysis and diagnosis process.

General preprocessing for images such as noise reduction and colour normalisation might be required to ensure that the important features that characterise the image are retained and not masked by unwanted artefacts. Normally, colour normalisation is carried out for research which utilises colour, as the staining concentrations are varied. Colour normalisation is generally a process to allow colour channel in image, loosen its dependency on the illumination, which is usually vary. A good colour normalisation techniques will produce a natural-look colour images which is not very different from the original images, but does not alter with illumination change.

Two reseach has reported the negative impact of not normalising colour in their images is in [11] and [109]. In [11], Bridges has used colour histogram to detect tumour in colorectal virtual slides and suggest a colour normalisation should be carried out as the detection accuracy has reduced when it is not. Then, Snape in [109] works with nuclei detection using colour deconvolution, but has skipped the colour normalisation process. The result is not promising, showing that features extracted from colour impacted factor should be colour normalised as well.

Magee et al. has reported a comparison of many colour normalisation methods on H&E and Diaminobenzidine (**DAB**) stained virtual slides in an attempt to solve the colour normalisation problem in histopathology images [71]. The work proposes colour deconvolution vector estimation for H&E and DAB stained images, thus suggesting that colour normalisation for histopathology slides are stain specific.

Noise reduction is also a very important process as all known feature extraction methods use the value of pixel, or group of pixels. The noise in an image can be minimised if the image is carefully (and manually) selected to ensure no artefacts or unnecessary background were included in the first place. Then, reducing the noise from images with series of image blurring, deblurring, thresholding and filtering will maximise the chance to pick up only the correct information.

These processes enable features measurements and manipulation on the pathology images without having to repeat any clinical setup or downgrading the original glass slides. However, not all extracted features are useful. Therefore, feature reduction and selection

might be needed to reduce the number of parameters to the most distinctive ones. Further testing is required to help in selecting key features that can be used to identify the characteristics of the objects in the image.

Then, using the selected key features in the previous stage, automated or semi-automated diagnosis of tissue disease can be made using (one or combination of) suitable machine learning techniques. The diagnosis is made by judging the tissue condition based on parameters measured from the image and inferences from the trained sample (if provided).

## 2.4 Tissue texture analysis

The most commonly used definitions to describe image texture are by Haralick et al [41] and Tamura et al [116]. Haralick first describes texture in an image as ‘the spatial distribution of tone variations within a band’ which can be evaluated as fine, coarse or smooth. Similarly, Tamura describes image texture as ‘a composition of coarseness, contrast, directionality, line likeness, regularity, and roughness’. It therefore, describe the arrangement of colour and intensities of the tissue image or certain detected structures in the tissue.

The advantage of using texture features is that the image texture can be extracted regardless of the shape and size of any pattern in the image. Figure 2.2 shows a sample of texture measurement, which is the irregularity (or randomness) values as described by Tamura et al. It is measured by the entropy of grey value of pixels in image. The randomness indicates how regular or repetitive a tone variation is over the whole image. However, the randomness value of the same image is different if it is scaled up or down (Figure 2.2 b, d and e), thus texture features are scale dependent.

There are two approaches to describe an image texture: syntactical and statistical approach [100]. Syntactical methods are not widely used, especially in medical applications [64]. They describes texture based on an analogy between the texture primitives spatial relations and the structure of a formal language. They also construct one grammar for each texture class during the training phase, and use this to determine the texture class, using varying method from adjacency probabilities, graph-like language, to tree grammar [79]. Basically, these methods are based on the idea of regular patterns exist in the texture. Therefore, we did not explore this idea as dysplasia in BO is a grey area.

Statistical approaches are widely applied, mainly because of the ability to measure the features and quantify the evaluation process. Properties or features which were extracted from images were analyse with statistical methods to understand the properties and patterns, as well to explore the data. However, there are two main texture feature

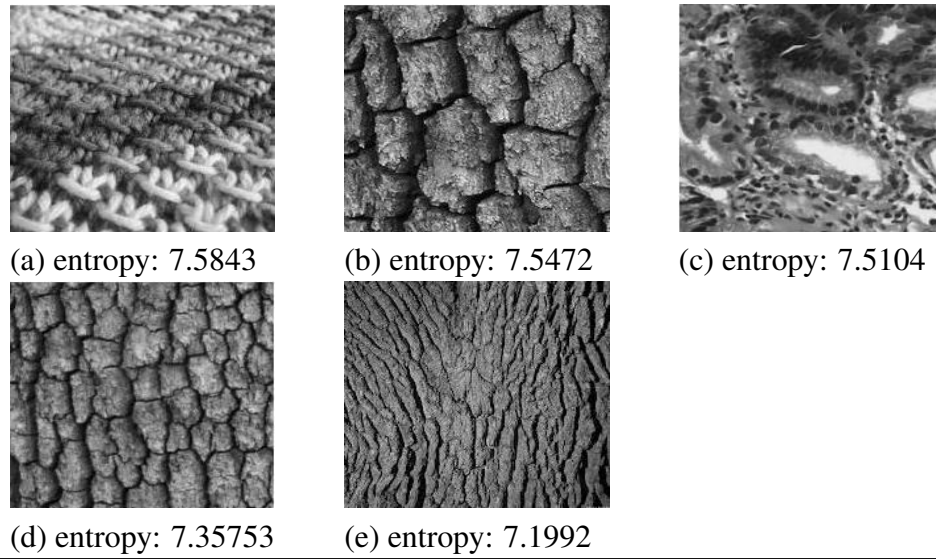


Figure 2.2: Sample of images with different texture. Note the entropy values for different images (a-c), but same image with different scaling produce higher difference (b,d,e).

extraction techniques, which are statistical and spectral. Both techniques have been used by researchers in medical imaging and have their own advantages and disadvantages.

Statistical feature extraction techniques use various sets of statistics from the distribution of image descriptors. The descriptors could be colours, intensities, edges, points, co-occurrence of pixels, patterns or any other local objects. Texture measures calculated from the original image are called first order measurements. Second order measurements consider the relationship between groups of two pixels in the original images whereas third and higher order consider the relationship among three or more pixels.

Spectral feature extraction techniques have to transform images into a spectral images using transformation models such as wavelet and fourier. The different frequencies of those signals or subsignals are measured as the image descriptors. With this method, the function to transform images and the sub-signals used as the thresholds are important features to represent image texture. Some published papers on medical image texture, implementing statistical and spectral measurement techniques will be discussed here.

### 2.4.1 Statistical feature extraction techniques

**Grey-level co-occurrence matrix (GLCM)** [41] is a second order statistical model, commonly used to describe texture features of an image. It actually counts the frequency of repeating pixel intensity (grey level) co-occurrence with its neighbouring pixels within a certain offset ( $\Delta$ ), in a certain directions ( $\Gamma$ ). Thus, the matrices indirectly contain the

image spatial relationships. GLCM also can represent the third order texture calculations when the considering the relationship among three or more neighbouring pixels.

Figure 2.4 illustrates the basic idea of GLCM, whereby four co-occurrences matrices are generated for a  $\Delta$  of one pixel (one immediate pixel), at four  $\Gamma$  ( $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  and  $135^\circ$ ). Therefore, the number of matrices generated depends on the  $\Delta$  and  $\Gamma$  value selected. However, the matrix size depends on number of bins selected to represent the image intensity. The example shown in Figure 2.4 uses three bins of image intensity [0,1,2].

These matrices are symmetrical where the pairwise relationship are counted in both opposite directions (eg:  $1 \Rightarrow 2$  and  $2 \Rightarrow 1$  are counted as two difference co-occurrences). To do this, the produced matrix is transposed and added to the original matrix. Thus, the relationship between pixel  $i$  and  $j$  is indistiguishable to  $j$  and  $i$ . Consequently, the total number of all element  $i,j$  in the whole image will become double. Sample as shown in Figure 2.3, the original GLCM matrix  $A$  is transposed into  $A'$ , and added to become matrix  $B$ .

$$\begin{pmatrix} W & X \\ Y & Z \end{pmatrix}_A + \begin{pmatrix} W & Y \\ X & Z \end{pmatrix}_{A'} = \begin{pmatrix} W+W & X+Y \\ Y+X & Z+Z \end{pmatrix}_B$$

Figure 2.3: Matrices operation to produce symmetrical matrix.

The symmetrical matrices are then normalised by dividing each element in the matrix with the sum of  $i,j$ . Therefore, the matrices will now contain the probabilities of finding the relationship of  $i, j$  in the image. From the generated matrices, texture features can be computed and the four most widely used GLCM textures are contrast, correlation, energy and homogeneity.

Contrast measures the intensity difference between a pixels and its neighbour over the whole image. It shows the sharpness of dark and bright pixels in image and calculated as (2.2a), thus a constant image (no intensity differences between pixels) would return 0. Whereas, correlation measures the degree of inter-relatedness between a pixel to its neighbour over the whole image. Values calculated with equation (2.2b) ranges between -1:1 where -1 means negatively correlated and 1 means positively correlated. Else, it will return 'Not a Number' (NaN) for constant image as the standard deviation ( $\sigma$ ) of the two pixels are zero (refer to Equation 2.2b).

Energy on the other hand is the change of pixel value in the image, which is the absolute difference of grey level value in the whole image as in equation (2.2c). Opposite to contrast, constant image will have the highest energy. Figure 2.6 demonstrates the differences of these values between two images with different texture.

Homogeneity measure the closeness of frequency distribution in GLCM to the GLCM diagonal as in Equation (2.2d). The closest value is one, which means the value of pixels are not varying much.

$$Contrast = \sum_{i,j} i, j^2 P(i, j) \quad (2.2a)$$

$$Correlation = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j)P(i, j)}{\sigma_i \sigma_j} \quad (2.2b)$$

$$Energy = \sum_{i,j} P(i, j)^2 \quad (2.2c)$$

$$Homogeneity = \sum_{i,j} \frac{P(i, j)}{1 + |i - j|} \quad (2.2d)$$

where  $N$  is the number of grey levels in image

$P_{i,j}$  the element  $i,j$  of the normalised matrix

$\mu$  is the mean intensity values in the matrix

$\sigma^2$  is the intensities variance

However, using grey images will eliminate any information that we might need in the saturation and hue channel. Similar application can be implemented in colour images as well, known as Multiple Channel Matrix (MCM) [92]. MCM counts the occurrence of pairwise values across different colour channel, for example red to green or green to blue. Therefore, for Red, Green and Blue (**RGB**) or Hue, Saturation and value (**HSV**) images, 6 matrices could be generated, as illustrated in Figure 2.5.

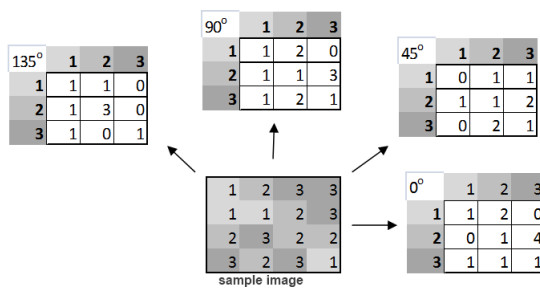


Figure 2.4: Asymmetrical GLCM generated from sample image.

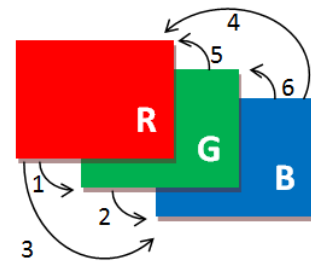


Figure 2.5: MCM channel pairwise in RGB images.

One of the implementations of textural features on colon virtual slides was reported in [40]. 82 textural features including GLCM, optical density (**OD**) histogram and grey level-run lengths were extracted for each image. After feature reduction, only energy, homogeneity, contrast and correlation of GLCM and OD were used to classify dysplas-

tic tissue with 83% classification accuracy. However, normal colon tissues are different from normal oesophagus tissue. This work also focuses only on locating dysplasia, not grading them. In addition, their research was closely supervised where each region is pre-classified. The tissue texture complexity may be simplified into only two obvious grades; normal and dysplastic. Thus linear classification rule can be used to locate dysplastic tissue correctly.

There were also other attempts to use GLCM and MCM texture features in colon tissue. Shuttleworth in [107] managed to classify normal, polyps and tumors in colon tissue using both, but demonstrated that using MCM can increase the classification result. The features investigated includes (but not limited to ) entropy, contrast, correlation, homogeneity, dissimilarity, angular second moments, energy, horizontal mean, vertical mean, horizontal variance, vertical variance as well as their standard deviation.

Another application in colon biopsy tissue is reported in [76, 77], where energy, inertia and homogeneity of the GLCM and morphological features were used. Using the linear discriminant analysis to reduce number of feature and support vector machine , 70%-100% classification accuracy achieved, however data samples were small.

GLCM and MCM applied on prostate tissue as reported in [22]; has achieved 80% AP for Grade 3 and 4 of gleason grading with KNN and Support Vector Machine (SVM). The GLCM feature used were energy, contrast, homogeneity, inverse difference moment, entropy, correlation, variance and information measure of correlation. The same features extracted for MCM, but between six combinations of channels in both RGB and HSV colour model. However, the grading performed well if both texture measurements were combined, but failed on their own. In addition, the result can only be achieved if the overall hue value of training and test set images are similar. Therefore, another attempt to grade prostate cancer according to Gleason Grading is reported in [57], achieving better result by implementing texton-based features on SVM and KNN.

The implementation of tissue texture features has also been carried out to analyse liver disease. Wang et al. reported in [123] that normal and diseased liver tissue can be differentiated using GLCM features on ultrasonography and computer tomography (CT) imaging on liver tissue. The GLCM fetures used was pattern, fineness, coarseness, correlation and dispersion in 4 directions. Susomboon et al. uses 9 Haralick's features of GLCM namely entropy, energy, contrast, sum average, variance, correlation, maximum probability, inverse difference moment and cluster tendency, as reported in [114]. these featurews were used on tomography images of liver tissue, to quantify the homogeneity and consistency of soft tissue in liver. Comphuwiset et al. on the on the hand used the nagular second moment, contrast, correlation, inverse of different moment and entropy



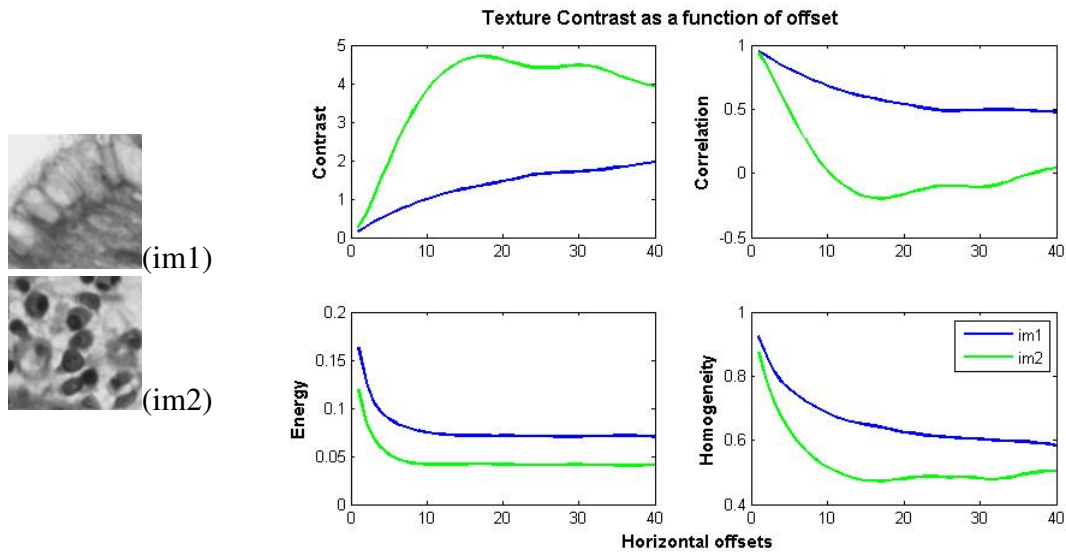


Figure 2.6: The different texture of epithelial surface compared to lamina region in oesophagus tissue. The four glcm features values are shown in the graphs provided.

as well as smoothness and uniformity of GLCM with statistical measures to locate and classify nuclei in attempt to detect glands in liver tissue biopsy images.

Implementation of texture feature extraction on breast tissue has also been published [50, 58, 74, 118]. Early investigation on texture features to classify benign and malignant tissue of mammogram images of breast is reported in [118] where it is concluded that the texture features show significant discriminatory power to differentiate the two classes. Martins et al. [74], Khuzi et al. [58] and Karahaliou et al. [50] investigate texture features on mammogram images as well.

In [74], the GLCM feature was reduced from 13 to 5 (namely contrast, homogeneity, inverse difference moment, entropy and energy) using feed forward selection method, and used in Bayesian Neural Network to classify the region to either normal, benign or malignant. The classifications were binary, and regions of interest were manually selected, hence a good classification result was achieved. Investigating texture features including 13 of GLCM (angular second moment, entropy, contrast, local homogeneity, correlation, shade, prominence, variance, sum average, sum entropy, contrast, difference entropy, sum variance and difference variance) on microcalcification of breast mammogram images, it is concluded that combination of GLCM and energy measure perform better than first-order statistics and grey level run length [50]. Khuzi et al. [58] has also demonstrated the ability of contrast, correlation, energy and homogeneity features of GLCM, to represent texture of normal and masses region in breast mammogram images.

The implementation of statistical measures in texture features extraction and imple-

mentation does not stop there. Other research was reported including application in lymphoma tissue [105]. In this report, a GLCM was created based on colour neighbourhood from colour quantized images on each selected region, rather than pixel intensities. Five GLCM features were selected: homogeneity, energy, contrast, correlation and entropy. Using a Bayesian classifier, images were classified into low, medium or high lymphoma and classification result for high grade lymphoma outperformed low and medium grade.

Yinhai Wang et al. [122] used tissue texture from GLCM and statistical moments for a supervised classification and segmentation of squamous epithelial layer from stroma and slide background. The segmentation has achieved 94.1% accuracy percentage at 2X, and 92.1% at 40X magnification, with SVM and pathological rules for error correction. The GLCM features used were contrast, correlation, angular second moment and inverse difference moment.

Felipe et al. and Kayser et al. in [26, 54] implemented texture features on various tissues of the human body. Felipe et al. used GLCM, and generate a smoothness measurement across image pixels. Known as gradient descriptors, it is calculated by summing up the number of occurrences in the GLCM for each  $|i-j|$ . The gradient vector was used as image signature by averaging the values obtained with the different distances for each direction. Different directions were implemented on each image to make it rotation invariant. During the image retrieval algorithm, the input image signature is compared with the queries images to find their matches. Kayser et al. [54] has also used similar approach on H&E stained images with different statistical measurement and machine learning techniques. Five features from texture (coarseness, contrast, directionality, line-likeness, regularity and roughness), object and architecture were used and 95% to 100% accuracy was achieved for crude classification.

Therefore, we can see extensive implementation of GLCM features, especially the contrast, correlation, energy and homogeneity in many domains and types of images. The basic concepts of pixel's value co-occurrence were retained, but the pixel's value itself were varied. Images and the pixel values can be intensity from different bit-depth grey level, or using colour value from many colour model (RGB, HSV etc), as well as colour quantized value. Results are mostly significant in capturing image texture.

**Grey level difference** on the other hand utilises the absolute difference of grey level value between pairs of pixels, or the differences between average grey level value [125]. 256 elements can be computed by taking the absolute differences of all possible pairs of grey levels distance at angle  $\theta$ , and counting the frequency. The frequency is normalized by dividing each element of vector by number of possible pairs. Other statistical measurements generated from the frequencies of grey level difference are the mean, entropy,

variance and contrast. Mean of grey level difference shows image coarseness, entropy shows the value of uniformity /complexity of image and variance shows the distribution of the grey level difference on the image.

Another method manipulating image grey level values in an image is **grey level run length**. This method use sets of linearly connected pixels belonging to same grey level [93]. The length of the run is the number of image points within the run. Therefore, images with coarse texture tend to have frequent long runs while finer texture dominated by short runs. Image texture is characterized by the grey level value, length of run (number of pixels in each run) and the run direction (usually 4, similar to GLCM).

A slightly different manipulation of image grey level values in extracting image texture features is **texton based**, where random patches are created across the whole images and transformed into another representation (usually using mathematical formula on a pixel's value) [93]. The representation of patches is used as a filter bank describing certain textures. Therefore, images are expected to have uniformity of patterns, or low noise over the texture [72]. In other words, the texton should only be applied within a texture class. Behrenbruch in [8] has applied a texton based approach on a classification of mammogram images into Wolfe Parenchymal Patterns [126]. The result is encouraging, however medical images are usually complex image with high noise or complex pattern. In addition, random sampling as implemented in the texton model might not find the relationship between the region and the neighbourhood area.

The implementation of textons has been extended into a **Bag of Words (BoW)** model, where the image is treated as a word as in document. The three main steps involved are feature detection, feature representations and codebook generation [25]. Therefore, collections of image features such as edges, intensities, pixels values and others are detected on random sampling of patches across the image. Then, these features are represented as numerical vectors and called feature descriptors.

A good descriptor should have the ability to handle intensity, rotation, scale and affine variations in the image. These descriptors are then converted into a 'codeword' (as analogy to words in documents), and 'codebook' (as analogy to a word dictionary) for the image. This is carried out by vector quantization, or clustering of descriptors. Thus, each patch is mapped into a certain codeword based on the feature similarity. Finally the image is represented by the histogram (or frequency) of the codewords in the image. Based on this technique, Caicedo et. al manage to characterize the distribution of local features of skin cancer histology images [13]. However, by taking the histogram of codewords, the distribution of spatial features in images are ignored.

The difference between bag of words and the texton based approach is the existence

6	5	2
7	<b>6</b>	1
9	3	7

(a)

1	0	0
1		0
1	0	1

(b)

1	2	4
8		16
32	64	128

(c)

1	0	0
8	<b>169</b>	0
32	0	128

(d)

Figure 2.7: Matrix operation to calculate LBP value

of the codebook (or the dictionary), used as a reference to cluster the patches. Usually, for one-class texton or feature descriptors problems, a dictionary or codebooks are generated by unsupervised learning throughout the whole training set [31]. This approach: generating codebook/dictionary with unsupervised classification, is known as **dictionary learning**. The implementation of dictionary learning in medical images has been published in 3D medical image denoising [65] and deformable image segmentation [133].

Other works implementing patches to sample texture in a smaller region can be seen in [132] and [21]. Yu in [132] has created patches of  $64 \times 64$  pixels on gastrointestinal (GI) tract histology images, and used Hidden Markov Model to find the correlation among them. The information was used to characterized digestive organs along the GI. Unlike Yu Diamond et. al. in [21], they analyze patches of prostate histology images individually and focus on the image character rather than the relationship between patches.

Another approach to represent image features is **Local Binary Patterns (LBP)** as introduced by [89]. With this approach, windows sized  $3 \times 3$  pixels are moved across images, to ‘localize the pixel value to its binary neighbours’ with a certain weight factor. The implementation is as illustrated in Figure 2.7 where Figure 2.7(a) is the original pixel value and the middle pixels value (denoted with bolded font cell) will be used as the threshold for creating its binary matrix, Figure 2.7(b). The matrix containing weight values in Figure 2.7(c) is multiplied with the binary matrix, yielding matrix Figure 2.7(d). The value of the middle pixel is calculated by summing up the numbers in Figure 2.7(d).

LBP is computationally simple, could be combined with contrast measures and is not sensitive to the grey scaling. Nevertheless, it might fail to represent types of image texture where image rotation and direction is worth investigating. Qureshi et al. in [96] has implemented LBP and wavelet subbands in their research on brain tissue. Principal Component Analysis (PCA), diffusion map and LDA were used to reduce the feature dimension and the selected features were sent to a Bayesian Classifier, KNN and SVM for comparison. Their wavelets features performs well compared to LBP.

Another approach to describe image texture feature are using patch/blocks/regions in image. One of the well-known techniques is the **Scale Invariant Feature Transform (SIFT)** [68, 69], where histogram of local position-dependent gradient orientation are

used as the descriptors. In SIFT, point of interest are computed from the differences between the adjacent levels in the Gaussian pyramids. The gaussian pyramid is constructed by repeatedly smoothing and subsampling of the input image. SIFT is invariant to scaling as local neighbourhood are normalised, and rotation invariant as it uses dominant orientation in the local neighbourhood. SIFT also contrast-invariance when the descriptors are normalised to unit sum, indirectly making it robust to illumination changes. However, the implementation of SIFT is best for matching or to locate specific structure from an image containing many other unwanted objects. The relative position of SIFT features are assumed to be static between images, thus it is not suitable pathological images.

A few implementations of SIFT on medical images have been reported in [66] for brain CT image retrieval, and has concluded that SIFT based method performs better than co-occurrence methods, especially on images with rotation and scaling issues. Another application is to improve registration time and complexity of microscopic sequence image registration [117], while Irshad et. al in [46] implemented SIFT to detect mitosis in microscopy images.

**Histogram of Orientation Gradient (HOG)** is another image descriptor closely related to SIFT. Dalal and Triggs [17] defined an image descriptor from a set of gradient orientation histograms; the histogram of angles and gradients vectors computed on a uniformly spaced cell (small connected regions). The accuracy were improved by measuring the intensity across a block and use this value to normalise all cells within the block. Thus, the descriptors are invariant to illumination or shadowing as well. However, the implementation need a region of interest, and HOG usually implemented with a sliding window over the image.

## 2.4.2 Spectral feature extraction techniques

The implementation of spectral measurement in measuring medical image texture features can be exploited in histopathology images as well. It is carried out by transforming the image domain into another domain with a transformation function such as the Fourier transform or Wavelets.

The Fourier transform is used to decompose an image into its sine and cosine components. The input image are in spatial domain, and the output are in the frequency or also known as Fourier domain. The basic application is the discrete Fourier transform where the frequencies are sampled, thus it contains only a set of sample which are just enough to represent the spatial domain image. The number of frequencies corresponds to number of pixels in the domain image, and decomposed into sinusoidal components.

Thus, analysing certain frequencies of the image indirectly gives information regarding the geometric structure in the spatial domain.

Unlike the Fourier transform, the **wavelet transform** is based on functions that are localised in the spatial and frequency domains. These functions (scaling and wavelet functions) are known as wavelets (small waves). Wavelets fulfill certain self similarity conditions (refinement condition), which decompose the original signal into different frequency levels, also known as sub-bands [80, 95] and [6]. Wavelets functions measure local variation of intensity in row, column and diagonal directions of an image matrix(s).

The implementation of wavelet feature extraction has been published for prostate tissue histopathology images. Kourosh and Hamid in [47] attempted to automatically classify the malignancy level of grade 2,3,4 and 5 of the Gleason Grading System for H&E stained biopsy tissue. They managed to achieve 97% accuracy using short support, orthogonality, symmetry and vanishing moment features from multiwavelet feature extraction on k-nearest neighbour classifier (**KNN**). However, the feature space is large thus grading duration is longer and suffers from unnecessary load. Then, using a smaller feature space with colour based and wavelet transform, the classification reduced to 81% accuracy as reported by Tabesh et. al in [115].

**Gabor filters** are a type of wavelet transform. It incorporates gabor functions; which are sinusoidal plane wave of some frequencies. The frequency, orientation and bandwidth are controlled by 4 parameters: gabor operator, orientation, central frequency and filter orientation [29]. It is sensitive to signal or energy differences, rotation and spatial frequencies, thus correct parameters will result in good segmentation or classification result. The application of gabor filters in medical imaging are mostly to Magnetic Resonance Imaging (**MRI**), mammogram, CT and ultrascan images because of the nature of medical images. Applications on a non-medical image texture are plenty, for example in [30, 72, 104] and [59].

Other research using statistical and spectral measurement in extracting textural features is reported in [77], and later implemented for spatial features [78]. Using hyperspectral imaging on colon biopsy tissue image, measurement of narrow wave-bands (0.01 micrometers) to wide-bands (0.4-2.4 micrometers) are possible. Using wavelets however, reveals that overhead occurs easily as too many features could be extracted, thus a feature selection method is crucial. On the other hand, as many features can be extracted, 2D bands classification have a comparable performance to 3D bands.

Tabesh et al. also reported in [115] the implementation of spectral measurement of texture features where wavelet decomposition was extended to different colour channels and these new wavelet bands are invariant to rotation. The implementation in prostate

tissue sections, to discriminate stroma, benign and cancer tissue can be found in [102]. Multiple spectral bands from images were extracted and classified with Gaussian Classifier. Similar application but in multiscale image analysis is reported in [124] where the features were used to describe chromatin texture in grading invasive breast cancer.

Implementation of spectral measurement, but with gabor filters can be found in [84] to segment myocardial boundaries in MRI images. Each dimension in a gabor filter can be leveraged to accurately extract the motion of the myocardium. By adjusting the spatial frequency domain of sinusoidal function in gabor filters, the myocardial boundaries can be detected and tissue displacement is recovered.

## 2.5 Tissue architecture analysis

Tissue architecture also known as tissue structure, can be defined collectively [23, 32, 36] as general organisation or patterns or arrangement with other cells and/or local structures within a tissue. This includes the spatial relationship with other structures in tissue. Thus, detecting certain local structures is usually carried out before tissue architecture can be analysed. As a result, tissue architecture is specific to the tissue biological, cytological and morphological condition.

Another term which is being used as a feature is topological (also called spatial) features. These features involve the compartmentalisation of tissue area, and the arrangement of local features within or around those compartments. Distance, or space between those structures (local or boundary of the compartments), arrangement and dispersion of the structure is the key for spatial features. This feature has recently becoming a popular research matter in histopathology image analysis, but it is disease-specific characteristics [37]. However, due to some similarity of its modelling and algorithm with architectural features, I group them together in this report.

The implementation also usually needs local structure detection like nuclei, cells, glandular, fat, blood vessels, lumens and others. Usually, local structure detection and mathematical formulation were needed to extract tissue architecture features. Quantified parameters for cytological changes in each local feature are important too as it can give detailed information on each tissue. However, parameters to be measured are still unclear. In addition, parameters in local features might be too detailed to represent the disease symptoms in tissue, thus creating the unnecessary information overflow, pattern distractions and later, a need for dimension.

There are researchers trying to find the cytological features in a tissue before classifying it based on the features detected. Some researchers tried to identify and extract

information from glands [85, 86], nuclei [15, 43, 67, 85, 102, 129], blood vessel [101] and squamous epithelium [122].

One example of a local structure analysis is automated detection of gland and nuclei structure from prostate and breast pathology slides which has been reported in [85, 86]. Relevant information relating to tissue at pixel level (first level), relationships between pixels (second level) and also domain specific information were implemented. A Bayesian Classifier was used in the first level to generate likelihood of lumen, cytoplasm or nuclei from the stained tissue. The lumen likelihood and its neighbour was then used as an initial starting mask to find round objects. Finally, domain specific information (such as specific arrangement and relationship between histological structures) was used to remove false positive boundaries, and similar process was applied to detect nuclei in cell.

Boundaries of local structures in tissue were also detected by using an active contour model; a framework based on utilising an energy function around neighbouring pixels of the object's boundary [51]. In this model, the sum of external and internal energy at the current configuration is minimised by using a gradient based approach. The model is also known as the 'snake algorithm' as it moves or slides along the contour, to fit the data. The external energy is at its minimum on the contour boundary, while the internal energy is minimum when the connected points is generally smooth. This model is robust to noise in image, but it need a good initial points as well as the shape it supposed to find.

One research work utilising the idea of snake was reported in [43], where research on stained microscopic oesophageal tissue were conducted to segment the cells using an improved snake algorithm. In this work, the contour points were restricted to move along radial directions, to detect the cell nucleus. This is carried out by introducing a growing energy, replacing the energy minimization used in the traditional method. The growing energy will attract the snake towards a broader boundary, whereas in the traditional snake algorithm, initial point should be near the real boundary.

Local struture analysis was also used to segment nuclei in oral cancer and complex tissue sections, as reported in [67] and [129], respectively. Xiaodong et al. uses marker-controlled watershed and combine mean shift and a Kalman Filter to segment the nuclei with 97.6% cell tracking accuracy [129]. Loukas in [67] however uses only PCA to give the histogram of components. Classifications were carried out using heuristic processing. Another local structure that researchers are interested in is blood vessels. Comparison between mean shift method and spectral method for detecting blood vessels was reported in [101].

Identifying local structures in tissue need specific values to represent them, such as shape, roundness, compactness, radius, colour and many more. The drawback of this



approach is lengthy development time and task constraints as many structures may appear in a tissue. Furthermore the structure varies based on organ or tissue type itself. In addition, each structure needs their own detection and segmentation algorithm as they have different characteristics from each other.

The most commonly used model to quantify architectural features are delauney triangulations [23, 55, 60], voronoi tessellations (**VT**) [23, 36], trees or graphs [23, 36, 37, 60], as they can efficiently represent spatial data and structural information in images.

**Voronoi Tesellations (VT)** segments area into blocks by creating medial lines between points of interest. Therefore it can be used to estimate the boundary, shape and area belonging to each point, like cells or glands. On the other hand, **delauney triangulation** connects each point to create triangles with maximum angle for each corner without any point trapped inside the circumcircle of any triangle. Thus it is more suitable for space related problem like estimating the crowiness of point of interest.

Unlike VT or delauney triangulation, **minimum spanning tree (MST)** connects points based on the weights or distance between each point. The final output of MST is a single tree with branches showing the lowest cost or total weights from the start point to the end point. **Neighbourhood Graph** on the other hand, consist of sets of nodes and edges. The interrelation between nodes and edges can be observed as the spatial or structural features. A basic diagram or illustration of delauney triangulation, voronoi tesellation, tree and graph is shown in Figure 2.8. These modelling diagrams can measure various structural features such as roundness factor, number of sides, area of polygons, average distance, distance to neighbouring cells, edge length and many more.

Depending on type of organs, some tissue structure changes drastically on the tissue surfaces while others do not. For BO, the changes hasvebeen explained at chapter 1.1 as well as in Table 2.2. Morphological changes affecting tissue surfaces are called ‘border effect’ [60]. These can be seen clearly in Figure 2.9 where the general profile of the tissue surfaces for both tissue samples are considered smooth but the surface membrane has shifted to become curvier and more prominent when dysplasia has reached grade 3. Therefore, region selection is very important and this is still lacking in tissue architecture analysis. In addition, [32, 60, 113] and [23] agree that the tissue boundary has a major impact in tissue architecture analysis.

Some applications of delauney triangulation, VT and graphs in mapping the architecture of tissue or structure in medical image are cited here. Keenan in [55] has reported work to grade cervical intraepithelial neoplasia (**CINN**) in cervical biopsy slides. In the attempts, all detected nuclei in epithelial cervical tissue were interconnected using delauney triangulation. The delauney triangulation mesh was then used to quantify the tis-

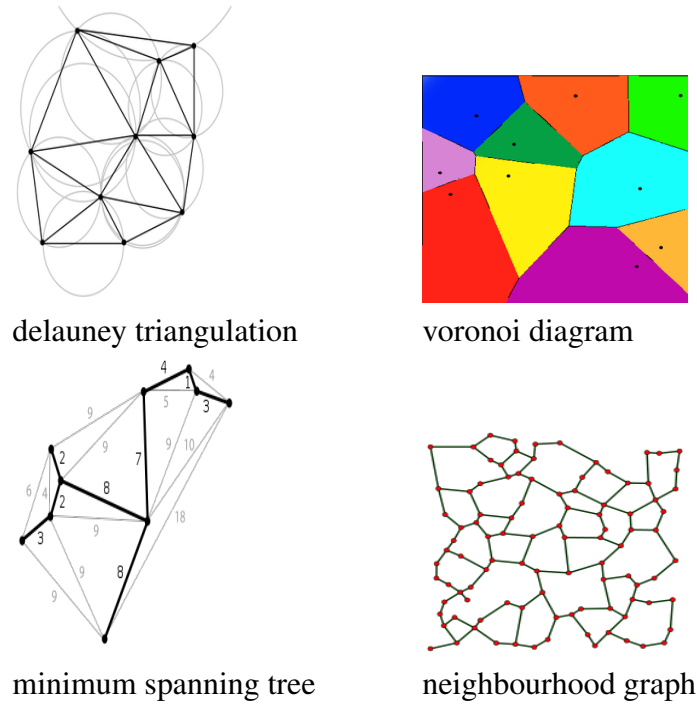


Figure 2.8: Sample of tissue architecture model

sue architecture with statistical analysis. Using discriminant analysis, 62.3% of the CINN cases were correctly graded with 0.502 and 0.415 KV for intraobserver and interobserver agreement respectively. However, this work faced a ‘border effect’ issue, where tissue regions containing the tissue border gave major impact on the tissue architecture. As a solution, the epithelial tissue images were digitally captured and carefully selected from cervical biopsies, thus providing a ‘clean’ region of epithelial layer without the tissue border for this research.

Another research implementing architecture features is by Landini et al. in [60]. This is carried out by detecting nuclei with colour deconvolution, and creating a cell boundary around the nuclei using the watershed transform method. Only cellular area in oral tissue analysis is mapped. As a result, a network graph from the local cell neighbourhood is obtained. Classification using discriminant analysis has achieved 67%, 100% and 80% accuracy in classifying normal, premalignant and malignant tissue. However, this work also suffers from ‘border effect’, thus images with tissue boundary were removed from the dataset.

Doyle et al. in [23] has attempted to automatically grade prostate cancer into 4 grades which are stroma, epithelium, grade 3 and grade 4. The work utilises a variety of feature analysis including architectural, textural and local structure features. Glands were

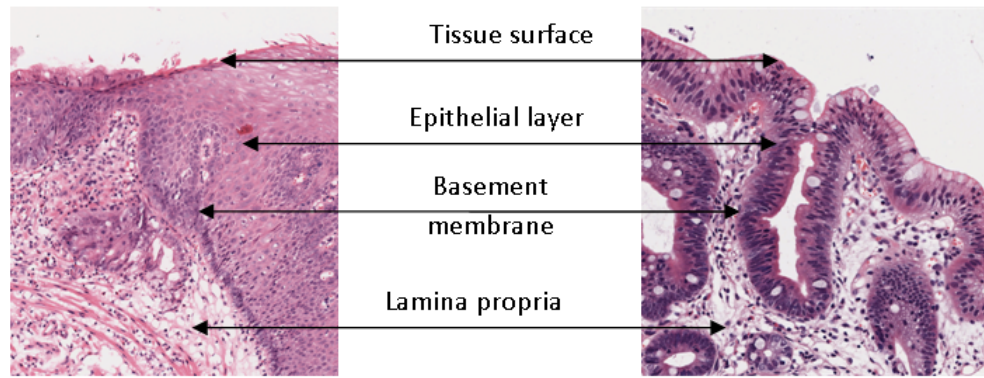


Figure 2.9: Sample of oesophagus tissue with tissue surfaces, basement membrane, epithelial layer and the lamina propria. (left) Sample of normal tissue with wider epithelial while (right) is sample of tissue with dysplasia grade 3. Narrower lamina layer and curvier profile of the basement membrane can be seen.

detected and used as points of interest for mapping the architectural feature using VD, delauney triangulation, MST and co-adjacency matrix. For textural features, first and second order statistical model were used together with gabor filters.

Grading was carried out on the basis of binary classification, with SVM. The best result to grade between grade 3 and grade 4 is (76.9%); using delauney triangulation features. However, grading between grade 3 and stroma gives the highest accuracy, which is 92.8% using average edge length of MST as the discriminant feature.

Grading prostate cancer has similar challenge to dysplasia in BO as it also has high variability of agreement and lack of standardisation. However, there is a known arrangement of structure, which is a group of 6 glands in prostate cell clustered together. Dysplasia in BO does not have such a luxury. In addition, this work also takes only tissue images without boundary, to avoid the ‘border effect’.

Tissue architecture analysis has also been implemented in breast tissue histology images. Basavanhally in [7] tried to map the arrangement of lymphocytes in breast tissue. A Bayesian Classifier and template matching was used to automatically segment lymphocytes. Using a manifold learning with graph based features from the detected lymphocytes nodes, 89.5% accuracy percentage of high grade breast cancer was achieved with SVM. Luckily, breast tissue histology images do not suffer from border effect.

Geusebroek et al. in [32] has tried to map only cellular areas on rat brain hippocampal tissue images. This was carried out by finding a correct architecture template from the images. To achieve this, cells were detected automatically, and connected with K-nearest graph and euclidean weighted graph. The average length from K-nearest graph gives a measurement of the tissue pattern scale which is normalized by the average of all distances

in the graphs. The graph is used as a template to reflect the tissue architecture.

Therefore, the difference between the selected template and the graph was obtained from the test image to determine a cellular area from the tissue. Geusebroek and the team managed to extract a flexible pattern of cell structure, but the approach used is only suitable for fields with regular patterns. In addition, this research also selected images without tissue boundaries in their study case.

It is anticipated that in tissue with complex morphological features/changes, many other local features also need to be considered to identify the tissue architecture. These may include goblet cells, columnar and squamous cell, crypts and ridges as well as glands.

An early attempt to grade dysplasia in BO was reported in [109], which has implemented texture and architecture features. All detected nuclei which were detected automatically are connected with delauney triangulation for architecture features while the textural features were extracted from colour histograms. The architecture features and texture features alone do not provide a good result but combined features however, managed to increase the average classification results to 62.5%.

## **2.6 Machine learning to enable pathologist-like diagnosis.**

In the previous subsections, several machine learning algorithms have been mentioned. Machine learning is an important phase in any artificial intelligence or pattern recognition system as it enables the computer to change its behaviour according to the data received. In tissue analysis, parameters that we have extracted in the feature extraction processes, will be used in a machine learning technique; to enable decision making from the given features. There are three types of learning; supervised, semi-supervised and unsupervised.

### **2.6.1 Supervised learning**

The supervised learning method is the most common machine learning implemented in medical fields. This approach enables machines to learn data patterns from samples containing data and the correct labels. During the learning process (also known as training), a supervised learning algorithms will learn to generalize patterns from samples and produced an inferred function to map the pattern into its label. These inferred functions will be used as a reference, or ‘knowledge’ for the machine to give a correct output when given a new set of data.

Sometimes, reasoning behind each selection or classification are understandable, therefore supervised learning is favourable in medical related systems. It can be implemented not only at the final decision or diagnosis level, but at feature classification level as well. However, depending on the training input and data size, overfitting may occur.

Overfitting is a condition where an algorithm memorizes patterns instead of learning and fail to generalise correctly from the training data [18]. This can happen if the features provided for training contain a lot of noise, or too many parameters compared to number of training data provided. Overfitting can cause learning algorithms to produce a rigid inference function which is not flexible enough to represent minor fluctuation of parameter changes.

**Bayesian learning** is an example of supervised learning. It uses the Bayes rules that assumes that the probability of elements belonging to any class is governed by a probability distribution. Decisions made by reasoning with the probability values and observing patterns in train data. The probabilities were changed many times throughout the learning process, and the final probability is used as the weight during the decision making process. The basic function used to calculate the probability of an event being in certain class  $P(y|x)$  is shown in Equation 2.3. However, the number of observed events should not be too small or the probability will be underestimated. Bayesian learning also enables the incorporation of prior knowledge through probability of each  $x$  and  $y$  or the certainty factor for each (or combination of) elements in the class.

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad (2.3)$$

**Gaussian Classifier (GC)** is another approach of supervised learning; used when the probability of the element being in each class was not available. Therefore, a discriminative approach is needed to model the  $P(y|x)$  or  $P(x|y)$  directly from samples. GC assumes that the probabilities are in normal (also known as gaussian) distributions, thus only the mean  $\mu$  and the standard deviation  $\sigma$  for each class and numerical attributes are needed. These values are used in Equation 2.4 to find the probability  $P(y|x)$ , or presented as 'belief' value in gaussian classifier. This equation is actually measuring the distance between elements based on the gaussian distribution model.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x-\mu}{2\sigma^2}} \quad (2.4)$$

There is also a supervised learning that can generate classification rules on certain conditions or structures, generally known as **decision tree (DT)**; where each leaf represents one classification rule. A decision tree will attempt to learn a pattern from the training

samples that maps from the attributes to the predicted class. A good tree is usually small as it is more easily understood and is likely to have higher predictive accuracy [94]. It starts from a root node, which is the rules that applies to others desicive nodes as well.

A divide and conquer tree method uses information gain and gain ratio. For example, if  $RF(C_j, S)$  is the function that denote the relative frequency of cases in  $S$  that belong to class  $C_j$ , the information content (or also known as entropy) that identifies the class for a case in  $S$  is as Equation 2.5.

$$I(S) = - \sum_{j=1}^t RF(C_j, S) \log(RF(C_j, S)) \quad (2.5)$$

where  $S$  is partitioned into  $S_1, S_2, S_3, \dots, S_t$ . The expected information requirement,  $G$  is then

$$G(S, B) = I(S) - \sum_{i=1}^t \frac{|S_i|}{|S|} I(S_i) \quad (2.6)$$

The gain criterion chooses the test  $B$  that maximise the gain ratio  $G(S, B)$ . The gain ratio  $P$  helps avoid overfitting with single, straight-forward cases by taking into account the potential information from the partition itself, and choose the test that maximize  $G(S, B)/P(S, B)$  among the test with least average gain. This is the proportion of information generated by the split  $P(S, B)$  as in equation (2.7) that is useful to create split desicions into nodes for classifications

$$P(S, B) = - \sum_{i=1}^t \frac{|S_i|}{|S|} \log\left(\frac{|S_i|}{|S|}\right) \quad (2.7)$$

Advantages of decision trees are the models are easily interpreted, regenerated, robust, usable in numerical and categorical data and may be incorporated with other machine learning techniques too. However, it is easily trapped into the overfitting problem as it works with greedy and recursively splitting algorithm. The prediction result usually come with high variance and sometimes with low accuracy. Therefore, **Random Forest (RF)** are introduced to increase the accuracy.

The random forest is an ensamble classification algorithm that uses DT as it basic classifier. The first steps is bagging; which is to create a forest of random sampling of learners which are independently trained on distinct bootstrap samples [10]. Bagging decreases test error by lowering the prediction variance with squared error loss, while leaving bias unchanged. The final prediction is the mean or class with maximum votes.

The **SVM** or Support Vector Machine as define earlier is another example of super-

vised learning, commonly used in medical field. It is a discriminative classifier to define an optimum separating hyperplanes between two classes in training samples. This is carried out by selecting a minimum number of critical boundary points in each class (also known as support vectors), and build the separating hyperplane with linear equation ( $f(x) = mx + c$ ) that separates the support vectors as widely as possible. Figure 2.10 illustrates the optimum hyperlane projected on a linearly separable classes.

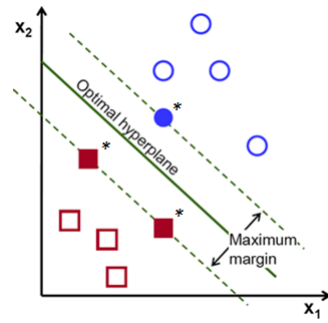


Figure 2.10: Optimum hyperlane projected by SVM learning algorithm for linear classification problem. The points denoted with (\*) is the support vectors, which is the closest points to the hyperlane.

The SVM has the advantage of modelling both linear and non-linearities data, resulting in complex mathematical models. The non-linear SVMs introduce a kernel function, which maps or transform the data points in a non-linear input space into a linear surface in feature space by function  $\Phi$  induced by kernel function  $k$  as illustrated in Figure 2.11. The  $\Phi$  and  $k$  are the key that enables the SVMs to find the non-linear decision functions with controllable complexity. The SVMs may also provide high accuracy percentage as compared to data mining [73]. However, the reasoning behind its classification is less understandable in a layman language. This can be a disadvantage in any automated diagnosis tool as it is very important to be able to explain clearly diagnostic results. Thus, SVM is used commonly for comparison or evaluation purposes.

The SVMs are commonly used in medical research for various types of implementation as in [7, 22, 23, 57, 96, 122] and [107]. Basavanthally et al. in [7] compared the use of SVMs with unsupervised Varma-Zisserman classification and reveals that SVMs outperformed the other by 60% to 90%. Doyle et al. in [23] managed to get 89.5% accuracy percentage in grading benign and cancerous cells. In [107], the SVMs were used to discriminate cancer and non-cancerous colon pathology slides while DiFranco in [22] used it to discriminate grade 3 and 4 of prostate cancer and Yin Hai et al. managed to get 92.1% accuracy in segmenting squamous epithelium from cervical pathology slides [122].

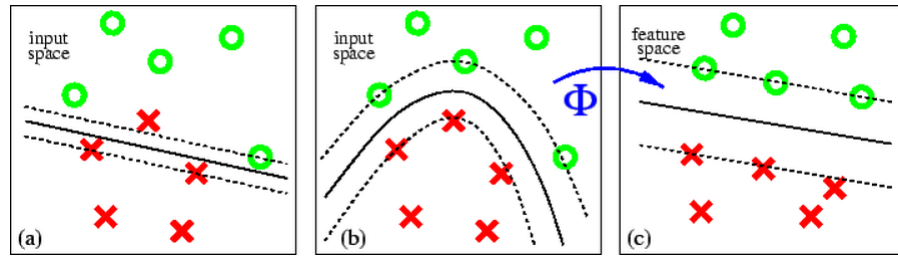


Figure 2.11: (a), (b), and (c), shows the same two class separation problem, but in three different view. Linear separation in for data points in input space is not possible with minimum errors (a) but a better separation is possible with a non-linear surface (b). This non-linear surface is in-fact a linear separation line when viewed from the feature space (c). (Image is courtesy from [81] )

### 2.6.2 Unsupervised learning

Unsupervised learning simply receives a sequence of inputs without desired output or rewards [33]. The approach of unsupervised learning is to observe patterns in the given data and use it to make its own classification or decision.

A commonly used unsupervised learning method is the **K-mean classifier, or k-means clustering**. The K-means classifier is a method to group features into  $k$  clusters based on the distance from the features to the cluster centroids. It justifies the grouping based on minimum distortion value, rather than fitting a probabilistic model. There are two key features: distance of each element from its cluster centroid, and the crowdedness of each cluster. Only parameters without a desired class, the distance functions and number of clusters are needed to use this classifier.

Another example of unsupervised learning is classifier based on **Principal Component Analysis (PCA)**. This classifier tries to find the best linear projection for each class separately. This is carried out by calculating the principal components of each class and designing the classifier accordingly during the training phase. The PCA subspace for each class learnt during the training phase is used as the class templates, and template-matching method is used for setting up the threshold. The classification of car and non-car, as reported in [127] was decided according to the distance of the new sample to these templates. PCA tries to find a new orthogonal coordinate system for features by incorporating the first few dimensions with the maximum variance of data [48]. The projection of the data is on the subspaces spanned by these principal components which correspond to different classes, thus it can learn from multiclass data. This approach is suitable for cases where the data of different classes are distributed in different style and different direction.

The principal axis line projection represents the strength of the data energy. Therefore, the energy distribution information can be viewed from different dimension, and even



enable user to project or eliminates some of them, which risking some information loss. Therefore PCA also used as a dimensionality reduction method. However, a good PCA application will still enable users to reconstruct the original data. The PCA Classifier results are unique and do not relate to any hypothesis about data probability. Therefore, prior knowledge cannot be incorporated.

Yiying et al in their paper [130] reported achieving 90% AP by using PCA Classifier to classify tumor tissues. Other researchers such as reported in [96, 102] and [67] have also decided to use PCA Classifiers in their work.

### 2.6.3 Semi-supervised learning

Semi-supervised learning sits between supervised and unsupervised learning, whereby patterns are learnt from a combination of label and unlabelled training data, which is independently identically distributed [134]. Its goal is to learn how combining both data may change the learning behavior, and design an algorithms that represents the learning behaviour.

A few assumptions have to be considered, which are smoothness, cluster and manifold assumptions [14]. Smoothness assumptions of supervised learning is to enable generalizing patterns from a training set but possibly covers the unseen test sets. This is done by assuming that two close points ( $x_1, x_2$ ) should correspond to output ( $y_1, y_2$ ). In other words, this assumptions refers to continuity, but termed as smoothness functions.

Cluster assumption assumed that points in the same cluster are likely to be of the same class. Depending on data population, the decision boundary for low density data separation should lie in a low-density region. For high density region, the decision boundary should cut a cluster into two different classes. Manifold assumption is dealing with 'curse of dimensionality' problem in many statistical methods. The facts that number of dimension causes volume to grows exponentially. Thus, semi-supervised learning assumed that the high dimensional data lie roughly on a low dimensional manifold. With this assumption, the learning algorithm can still operate in a space of corresponding dimension.

Semi-supervised learning is getting acceptance in many machine learning research nowadays as it can use readily available unlabelled data to improve supervised learning tasks, especially when the labelled data are scarce or expensive. The application of semi-supervised learning in medical image analysis have been published to classify pattern in MR brain images [28]. In this report, semi-supervised SVM has been used and gives high accuracy compared to traditional SVM. Another implementation is on high resolution computed tomography lung images to segment tissues in lung [108]. Combining k-means

clustering (unsupervised) and knowledge acquisition, existing class can be refined and gives high classification accuracy, similar to supervised learning.

## **2.7 Conclusions**

The volume of research papers in computer aided diagnosis is very encouraging since the existence of digital imaging techniques as presented in 2.2 and 2.3. Most established usage however stops at imaging technology while the diagnosis part is still to be incorporated into medical practice. Nevertheless, the usage has been used widely in research and education. Researchers are trying not only the supervised machine learning techniques in computerized diagnosis, but also unsupervised and semi-supervised learning to achieve better understanding, as well as making use of all available data.

Grading dysplasia is a well known issue among pathologists, and virtual slide has been actively researched and used in its education. There are common rules from consultant pathologists on how to decide a grade for dysplasia, but it is still appreciated that some rules disagrees with the others. It takes years of experience of diagnosing dysplasia in BO, to make a pathologist really confident of the grade they are giving, and to score a very good agreement for their intraobserver variations, as explained in 2.1.

Lack of prior knowledge which could be used as a ‘rule of thumb’ in supervised machine learning, and also long list of cytological features to observed as in Table 2.2 have motivated us to use texture features in discriminating between grades of dysplasia, rather than architecture. In addition, the nature of BO virtual slides tissue samples suffer alot from border effects, as explained in Chapter 2.5.

The ground truth images of BO virtual slides are very ambiguous in the way that in one virtual slide, we can have many annotated regions with different grades, and in each annotated region, lower grade of dysplasia may co-exist. A detailed explanation follows in Chapter 3.

## Chapter 3

### Tissue detection and selection process

---

A long list of cytological features to be observed (or to detected) as in Table 2.2 has motivated the use of texture features, instead of architectural features to discriminate between grades of dysplasia. However, as the important features lie in the tissue surface membrane, avoiding this area might cause major information loss.

In addition, only vague prior knowledge is available, which cannot be used as a 'rule of thumb'. No established map of normal oesophagus tissue texture or architecture is available. In addition, the ground truth images of BO pathology slides are very ambiguous in nature, where in one virtual slide we can have many annotated regions with different grades, and in each annotated region, multiple grades of dysplasia may co-exist, as shown in Figure 3.1.

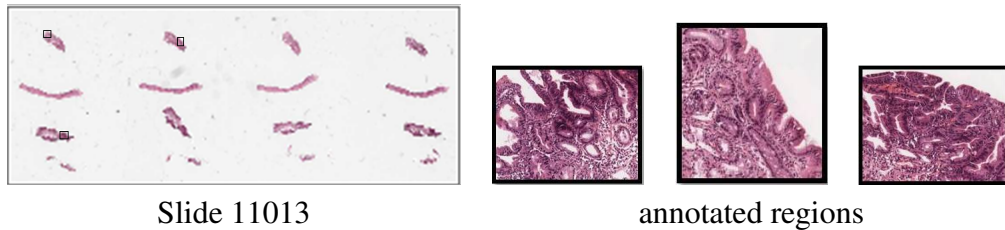


Figure 3.1: Slide 11013 with consensus grading as grade 4 but regions annotated marked as grade 4, grade 2 and grade 4 respectively.

There are two types of ground truth available for this research. The first is for the whole virtual slides. It contains the consensus diagnosis of BO for the whole slides, which naturally contain several biopsies with multiple grades of dysplasia. These sample

tissues are usually tissues that could be normal, or dysplasia with many grades as well as other abnormalities found. The second type of ground truth is the annotated region of the virtual slides, showing specific regions with dysplasia and the respective grade, based on the consultant pathologist's opinion.

This chapter contains an explanation for both types of our ground truth images, followed by the processes that have been carried out for image pre-processing such as tissue detection and location, artefacts elimination and noise removal. Then, two methods to create regions from each detected tissue are carried out with experiments, and compared. Finally, an existing colour normalisation technique is tested on images to help us choose between processing with greyscale or colour normalised images. This is important as the final images will be used in the next step: tissue texture analysis

### **3.1 Ground truth: Whole virtual slides**

One hundred and forty-eight H&E stained glass pathology slides of 127 patients from Leeds General Infirmary were selected by a consultant pathologist, Dr Darren Treanor; to make sure that each patient has evidence of BO with specialised intestinal metaplasia in at least one biopsy. These slides were also examined for the staining quality as some were faded. Where necessary, the original slides were de-cover slipped and re-stained, or re-cut from the block. Then, all slides were anonymised and given a random number as identification. The details of glass and virtual slides preparation as well as the pathologist involved can be found in [119].

These slides were sent to six UK national expert gastrointestinal pathologists for BO for dysplasia grading, following the categories shown in the far right column in Table 3.1. The grades given by them were independent, and solely based on morphological appearances of the tissue on a single slide.

For these selected glass slides, the overall interobserver agreement among the six expert GI pathologist scored for six grades of dysplasia following the Vienna classification (Table 3.1, the last column) was 57% with 0.41 KV. The grading were re-evaluated with binary classification in three broader classes, as shown in table 3.1 (the second last column), and higher agreement was achieved with 0.73 KV for G1, 0.37 KV for G3 and 0.73 KV for G5. The calculation for KV is as explained in Chapter 2.1.

Due to broken or cracked slides or thick cover slip of the glass slides, only 144 slides were scanned with Aperio T3 using a 40X objective lens to produce virtual slides with a final resolution of 0.23 microns per pixel. Images are stored in an Aperio server at


Severity changes increase	Modified classification(2 groups)	Modified classification: 3 groups	Vienna Classification of Dysplasia
	Non-dysplasia	G1	1-Barrett's only
			2-Atypia, probably negative for dysplasia
	Dysplasia	G3	3-Atypia, probably positive for dysplasia
			4-Low grade dysplasia (LGD)
		G5	5-High grade dysplasia (HGD)
			6-Intramucosal carcinoma (IMC)

Table 3.1: Range of dysplasia grading, which could be used.

St James's Hospital, Leeds at: <sup>1</sup>. These virtual slides were sent to two UK expert GI pathologists for dysplasia grading, independently. The experts were asked to annotate or mark the areas of tissue with prominent signs of dysplasia in each grade. At the end of the data collection time, diagnoses for only 140 virtual slides of this dataset were available.

Random selection of virtual slides for training and testing data in each grade was carried out, making sure that no region from the virtual slides test set was included in the training set in any other phase of modelling or training. Figure 3.2 illustrates the relationship between annotated images, virtual slides, and the training and test set for both annotated regions and virtual slides.

As the number of ground truth images for virtual slides and annotated images are limited for grade 3, we decided to follow the simpler grading suggestion by [56, 83]; which groups together Grade 1 and Grade 2 as G1, Grade 3 and Grade 4 as G3, and Grade 5 and Grade 6 as G5. This has been approved by our domain expert. Furthermore, the KV for three grades is higher than for six grades, showing the possibility of shared criteria between neighbouring grades. The number of available virtual slides according to the grade distribution, as well as the number of training and test images, is shown in Table 3.2 below.

<sup>1</sup>[http://slides.virtualpathology.leeds.ac.uk/Research\\_1/Darren/Barretts/](http://slides.virtualpathology.leeds.ac.uk/Research_1/Darren/Barretts/).

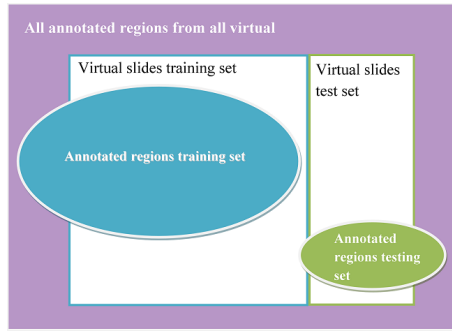


Figure 3.2: Relationship of virtual slides and annotated region training and test set.

Grade	Available	Train	Test
G1	81	20	5
G3	30	20	5
G5	29	20	5
total	140	60	15

Table 3.2: Available virtual slides, training and testing set according to grades.

Based on the amount of data for each grade, only 20 training and five test virtual slides images were used for each grade (G1, G3 and G5). This is to balance the number of data in each grade because G3 only contains thirty images. For these selected virtual slides, the interobserver agreement between the two consultant pathologists for six grades was 0.68 KV and increased to 0.77 KV for three grades (both are interpreted as good agreement), showing that our ground truth has a strong confidence level on each grading.

## 3.2 Ground truth: Annotated regions

From 140 virtual slides, both of the consultant pathologists marked 438 regions in total, to show signs of dysplasia in different stages and graded them accordingly. These annotated regions are at different magnification levels and sizes, while the location mark can be totally different or partially overlapped as each pathologist has examined and marked them independently.

Information about the annotated regions was recorded. This includes its virtual slide ID, region ID, coordinates, grade, comments by pathologist if any, as well as the magnification level used when the annotation were made. Therefore, the region can be re-extracted.

Oesophagus biopsy samples contain two major regions of oesophagus tissue which are the epithelium lining and the lamina propria (as in Figure 2.9). Annotated images as selected regions are very important to show the differences in tissue structure at each stage of dysplasia.

Each annotated region was examined together with a consultant pathologist (Dr. Darren Treanor) for technical and contextual quality and 22 images were excluded in the first filtration phase, as detailed in Table 3.3.

In the second filtration phase, annotated regions containing both the epithelium tis-

Reason	Number of region	regionID
Misleading annotation	15	621, 601, 576, 571, 367, 366, 333, 332, 331, 298, 279, 276, 275, 48, 522
Not H&E staining	2	577, 758
Size 0 pixel	1	30
Slide quality (over-lapped tissue)	2	111, 112
Abnormal staining	1	575
Blurred tissue	2	192, 244

Table 3.3: Excluded annotated regions.

sue and lamina propria are selected as [38, 75, 87] have reported that surface membrane (the epithelium lining) is an important feature in characterising dysplasia. Then, regions annotated from the validation set (as shown in Table 5.1) were removed as well. This is important to ensure that our validation set is valid and is not involved in the training process. Lastly, numbers of data for each grade were balanced, and as a result, 30 training and eight validation sets of annotated regions were selected at random.

Annotated images were varied in size but all were greater than 250 pixels in width  $\times$  250 pixels in height, and smaller than 5000 pixels width  $\times$  5000 pixels height. All annotated images were reviewed by a consultant pathologist and did not have any significant identifiable artefacts in them.

### 3.3 Noise reduction

Virtual slides and annotated images have a similar pre-processing which is artefact elimination, colour normalisation, boundary detection and region creation. However, they are processed at different magnification levels.

For virtual slides, images are recalled at thumbnail-sized and converted to a binary image. This is carried out by retaining pixels with an intensity of zero, while pixels with other values are replaced by 1, producing only black and white images. Pixels with zero value are considered as background. Based on eight connected pixels, the connected pixels were grouped together as objects.

Among these objects, there are artefacts and tissue samples which are not useful for the diagnosis, as shown in red squares in Figures 3.4(a). These can be some part of tissues which were torn from the main samples, smeared stains, blood or air bubbles trapped

between the glass and the cover slip. To filter these out, five virtual slides which contain 84 biopsy samples in total were used to set the threshold value in determining the optimum candidate tissue size from the detected objects. In the experiment, objects sized  $(\tau) < 100, 150, 200, 250 \text{ pixels}^2$  were eliminated. The number of false positive objects detected for each  $\tau$  was recorded as shown in the graph labelled Figure 3.3.  $\tau = 200 \text{ pixels}^2$  produced the minimum false positive detection rate (9) and therefore it is used as the threshold value. An example of filtered artefacts from a labelled image is as in Figure 3.4(b). Then, the remaining objects were accepted as candidate tissue and labelled, as in Figure 3.4(c).

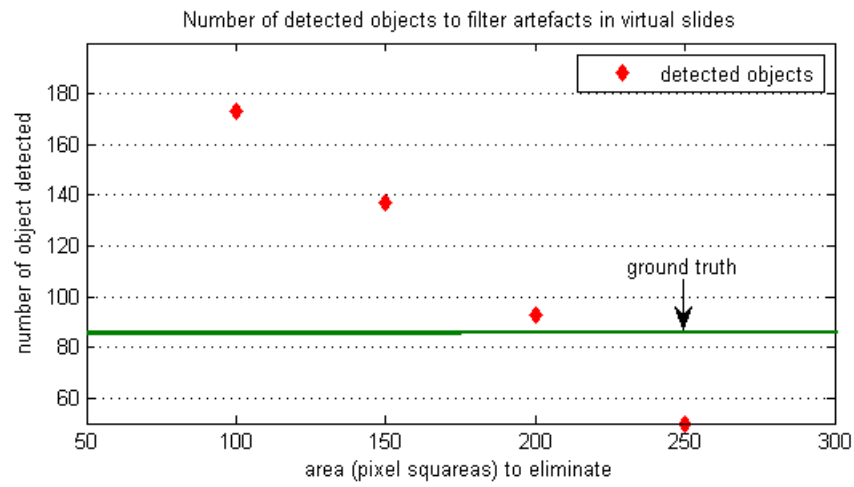


Figure 3.3: Graph plot showing number of detected objects with different values of eliminated area ( $\tau$ ).

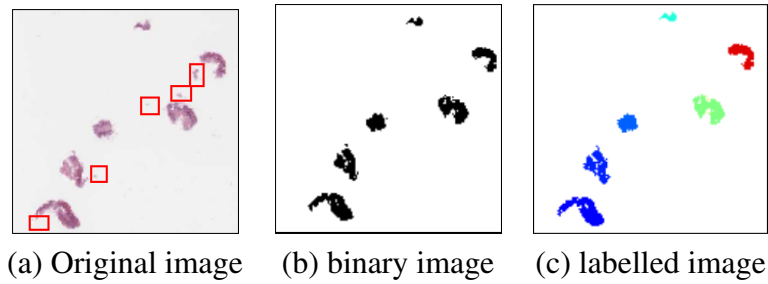


Figure 3.4: Image de-noising process



### 3.4 Region creation

Every candidate tissue in the whole virtual slides was divided into regions for further analysis. The objectives were to analyse individual tissue on a regional level and to learn the texture based on the ground truth of annotated regions. We tested two approaches to create the region: the first is automated region creation based on the complexity of the boundary curvatures, and the second is simple tile creation across the tissue image. These regions are filtered through a series of threshold values to ensure that they contain tissues that are big enough to include sufficient dysplastic cells images and its neighbouring areas

#### 3.4.1 Curvature-based regions

Our first approach is based on the fact that normal oesophagus tissues are lined with stratified squamous epithelium cells. The smooth lining eventually changes to form a villiform structure when dysplasia becoming more severe, as explained in 1.2 and shown in Figure 1.1. Therefore, we use the number of high curvature points along the tissue boundary to measure the complexity.

In order to do this, each candidate tissue is zoomed at 4X magnification in its binary form and a bounding box for each of them is created using its minimum and maximum coordinates of axis-x and axis-y. The coordinates of the intersection points between the candidate tissue and its bounding box, as well as all connected pixels with value '1', excluding the bounding box itself, are recorded. These recorded coordinates represent the candidate tissue boundary.

The shape of the tissue's boundary is analysed to get a general idea of the tissue condition. Therefore, window sizes ( $j$ ) [10, 20, 50, 100] were tested to identify the best representation of the tissue shape. Samples of candidate tissue boundary detected with eight connected pixels and different values of  $j$  is as shown in Figure 3.5.

In this phase, the curvatures of the candidate tissue boundary are used to discriminate the villiform patterns from the smooth boundary of normal squamous cells. An appropriate window size is needed to calculate and identify high curvature points ( $hcp$ ) along the tissue boundary without losing too much information.  $hcp$  are calculated using the coordinates of three adjacent points  $j$ , as in Equation 3.1.

$$curvature(rad) = \frac{[1 + ((\frac{dy}{dx})^2)^{\frac{3}{2}}]}{|\frac{d^2y}{dx^2}|} \quad (3.1)$$

where

$$\frac{d^2y}{dx^2} \approx \frac{\Delta m}{\Delta x} \text{ and}$$

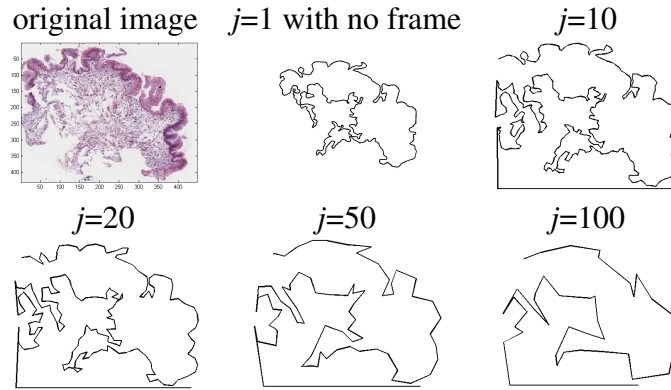


Figure 3.5: Image showing the tissue shape representative with different window size  $j$ .  $j=20$  is selected and the pixel along the image bounding box is eliminated.

$$\Delta m = \frac{\Delta y_{j,j-1}}{\Delta x_{j,j-1}} - \frac{\Delta y_{j+1,j}}{\Delta x_{j+1,j}}$$

To change radian into degree:

$$\theta = 180 \times \frac{(\text{acos}(\text{rad}))}{\pi}$$

There are two stages in identifying *hcp*. The first stage is to retrieve the whole tissue at 4X magnification and the *hcp* is used to identify the tissue regions. Then in the second stage, each region is recalled at 10X magnification, and the boundary complexity is refined again with new *hcp*.

At the first stage, a series of highly curved points are used to segment tissues into regions. The curvatures  $\theta$  between each point of  $j$  were calculated to determine the high curvature points *hcp*. *hcp* were defined by the change of  $\theta$  within a specific range  $[\theta_1, \theta_2]$ . The distances between each *hcp*,  $\omega$  and the standard deviation  $\omega_\sigma$  determine the regions in the tissue, where it starts from  $j$  with highest distance  $\omega_h$  to  $\omega_h - \omega_\sigma$ . Figure 3.6 (a) and (b) below illustrates these processes.

Referring to the illustration in Figure 3.6 (a), (a-d) are the points where the tissue boundary will be segmented, as the distances fall between highest distance  $\omega_h$  to  $\omega_h - \omega_\sigma$ . The first patch is from  $a$  to  $b$ , then from  $b$  to  $c$ , from  $c$  to  $d$  and lastly from  $d$  to  $a$ .

A brute-force approach, consisting of 1628 experiments with varying values for each parameter was carried out to fine-tune the best combination of  $j$ ,  $\theta_1, \theta_2$  and  $\omega_t$ . *hcp* detected were compared to a hand-marked images for evaluation. The average precision and recall resulting from five-fold validation was used to set the threshold values for each parameter.

At 4X magnification level, it was found that  $j$  of 20 pixels best represented the tissue peaks and crypts and the  $\theta$  between  $30^\circ$  and  $150^\circ$  was selected as a threshold to define a

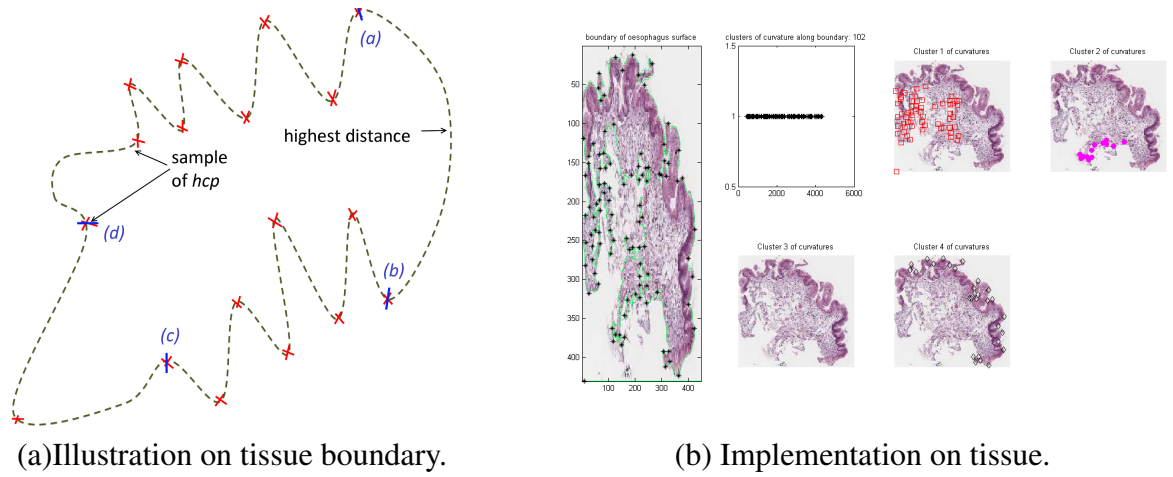


Figure 3.6: Implementation of region creation.

$hcp$ , with the  $(\omega_t)$  is 200 pixels. This parameter setting produces the minimum average false positive curves (5.87 per 4091 pixels) in the tissue boundary when compared to manually marked up peaks. Thus, it has minimised the sensitivity of the curves detection algorithm as the sensitivity is not needed at this stage and is the reason why false negatives are counted for decision making.

At 10X magnification, the best average  $j$  is 40 and  $\theta$  between  $40^\circ$  and  $160^\circ$ . A threshold of 30  $hcps$  per 100 pixels along the epithelial layer to class a boundary as ‘complex’, was established using 40 training patches annotated by a pathologist. Tissues are segmented into regions using the first and last  $hcp$  of the boundary classification. The classification is based on threshold value for the number of high curvature points detected per 100 pixels along the boundary. Classification samples of smooth and complex boundaries are shown in Figure 3.7, and the results for both stages are simplified in Table 3.4.

zoom	$j$	$\theta_1$	$\theta_2$	threshold	Precision	Recall
4	20	$30^\circ$	$150^\circ$	$(\omega_t)=200$	93.83%	90.61%
10	40	$40^\circ$	$160^\circ$	complexity=30 $hcp/100pix$	81.40%	72.92%

Table 3.4: Parameter setting for complexity measurement.

This method enables automatic region selection to analyse tissue in virtual slides. By knowing the most likely area of dysplasia (the complex boundary), we can focus on these regions for the next analysis steps, reducing the time and overhead in region-level analysis process. However, this method caused a repeating analysis of the same internal tissue area which affected the whole tissue analysis process. Examples of regions created with this method are highlighted in Figure 3.8. Therefore, the second approach is applied where

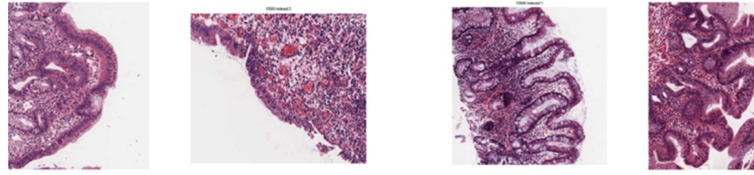


Figure 3.7: Tissue classification based on *hcp* thresholding. The first two images from left are classified as smooth, and the last two are complex.

the tissue is divided into tiles of regions.

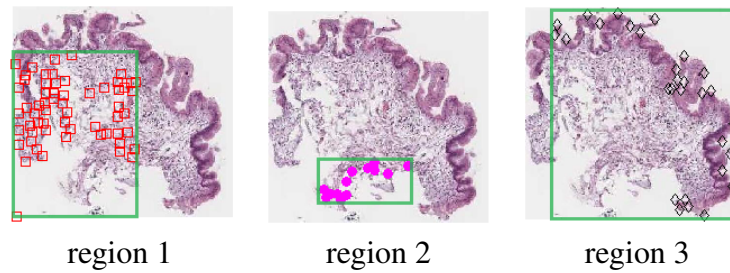


Figure 3.8: Tissue regions created based on complex and smooth boundaries from *hcp*. Note the overlapping areas of region 1 and 2 in region 3.

### 3.4.2 Tile-based regions

The second approach is to divide tissues into regions by cutting tissues into tiles of maximum size between  $1800 \times 1800$  pixels width and height due to software constraints. This approach is simpler to compute and it does not look at the overall tissue shapes. The main objective is to include as much tissue content as possible into a region for texture analysis. Therefore, tissues are retrieved at 40X magnification. This can also overcome cracked image issues that were caused by incompatible resolutions. At this point, another level of filtration process is used to ensure only meaningful regions are passed through the image analysis process.

In order to divide images into tiles, the size of the tissue's bounding box is retrieved first. As texture features rely on pixel values, region creation was carried out based on the number of pixels in a region. Therefore, the tissue's width is divided into 1800 pixels to get the number of columns available for the tissue. A similar process is used for the tissue's height, to calculate the number of rows. Then the tiles of the region are indexed or numbered accordingly for tissue image reconstruction later. The smallest region accepted is 800 pixels wide  $\times$  800 pixels high in order to ensure that we have enough tissues in the region.

The grey level values for these candidate regions are stored for artefact elimination. Based on our experiments, images containing tissues are those where the average grey level is  $<0.73$  or the entropy is  $<0.73$  or the entropy is  $>6.3$ . Wax smears (as shown in Figure 3.9) can be eliminated using an additional condition where the sum of grey level histogram between bin 190 to 210 is higher than the overall mean histogram. These values are sampled from 15 randomly selected artefacts discovered while processing the virtual slides. Some examples of the successfully eliminated artefacts are shown in Figure 3.9.

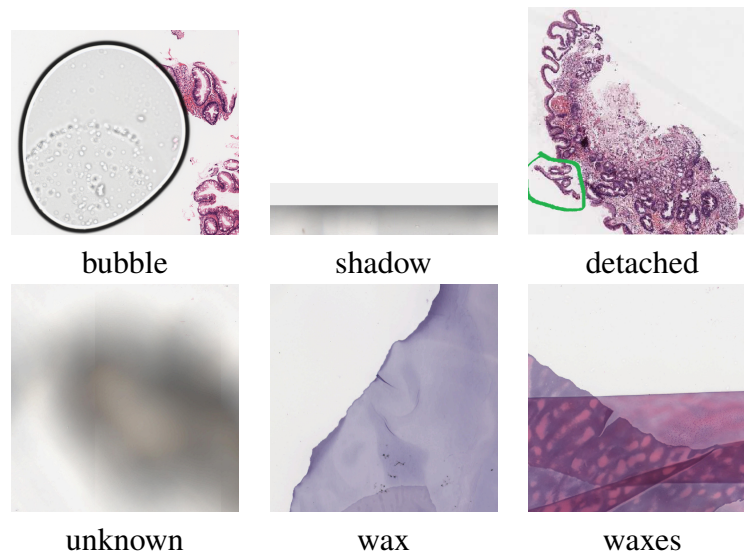


Figure 3.9: Non-tissue artefacts commonly found in virtual pathology slides.

To make sure that we only capture regions containing oesophagus tissue with its epithelial tissue, only tiles containing both tissue and background values are selected. However, some controlled value or a threshold value, is needed to control the possibility of having too few tissue samples (or too much background included) in a region. Additional criteria are enforced to help select only the meaningful regions. Based on experiments on these 84 images, regions where the average grey level values are  $\geq 0.75$  and total white pixel (background) is  $\leq 35\%$  are selected. Figure 3.10 illustrates our second method and the omitted regions are highlighted.

### 3.5 Colour Normalisation

Tissues in virtual slides have a variety of staining concentration, thus affecting the colour. Therefore, colour normalisation as explained in Chapter 2.3 was carried out on a set of test images to test whether or not we should use colour or greyscale images. Seven anno-

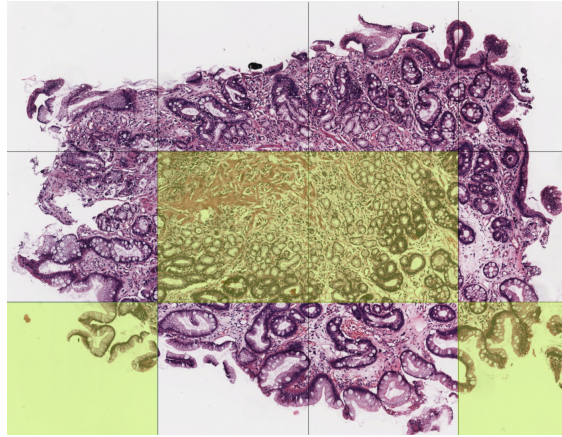


Figure 3.10: Second method of region creation, with selected region.

tated images were selected for each grade, including samples of tissue with and without epithelial membrane. The normalisation model proposed by Magee et al. in [71] is tested on our images as the code and paper are ready for quick setup.

### 3.5.1 Normalisation method

Seven annotated images of BO virtual slides were selected based on their colour variations for testing. These test images varied in size as well, but standardised magnification was to 20X. The test images in Table 3.5 show variations of colour saturation and size. Each image is normalised using the code and relevance vector model provided along with [71].

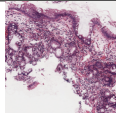
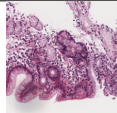
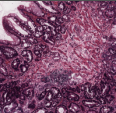
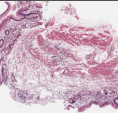
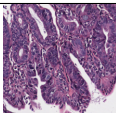
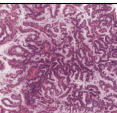
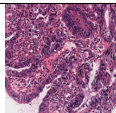
annID	7	55	158	290
image				
size(pix)	1908*1984	976*956	5056*4440	7208*5816
annID	427	618	612	
image				
size(pix)	724*716	7328*5800	1834*1734	

Table 3.5: Images used to test combinations of different normalisation techniques with different colour classifier models.

Magee et al. in their paper [71] evaluated five colour normalisation methods which map images into their target image either with pixel class or stain-specific colour deconvolution representation as explained as below:

- Reinhard

Reinhard et al. present a linear colour transformation in an  $l\alpha\beta$  colour space [98]. In this model, the colour distribution of an image is transformed to match the colour distribution of its target image. Thus, the means and standard deviations of each channel in both images are the same. The transformation, as shown in Equation 3.2 is applied at each pixel where classes are assigned based on the largest probability. However, this method is better performed on well segmented classes of pixels (e.g. sea, beach etc.) and might not work well on a blended image like H&E stained virtual slides.

$$l_{mapped} = \frac{l_{original} - \bar{l}_{original}}{\hat{l}_{original}} \hat{l}_{target} + \bar{l}_{target} \quad (3.2a)$$

$$\alpha_{mapped} = \frac{\alpha_{original} - \bar{\alpha}_{original}}{\hat{\alpha}_{original}} \hat{\alpha}_{target} + \bar{\alpha}_{target} \quad (3.2b)$$

$$\beta_{mapped} = \frac{\beta_{original} - \bar{\beta}_{original}}{\hat{\beta}_{original}} \hat{\beta}_{target} + \bar{\beta}_{target} \quad (3.2c)$$

where  $\bar{l}$ ,  $\bar{\alpha}$  and  $\bar{\beta}$  are the channel means and  $\hat{l}$ ,  $\hat{\alpha}$  and  $\hat{\beta}$  are the channels standard deviations over all pixels in the image.

- VB Reinhard Hard

This approach is suggested by Reinhard himself where pixel class should be known prior to the transformation. Thus, Magee et. al. in [71] used a probabilistic prior and a Variational-Bayesian Gaussian Mixture model to estimate the colour of pixels belonging to hematoxylin, eosin stains or the background.

- VB Reinhard Weighted

This is a further extension of the previous VB ReinhardHard, where the colour estimation, standard deviation and transform calculation are weighted.

- CDV Linear

This model uses a linear transformation model on RGB channel and sees that each channel matches with the mean and standard deviation respectively.

- CDV Multimodal

This approach uses two component 1D Gaussian mixture models in each RGB channel to separate background pixels and foreground.



In the first three normalisation methods, ten relevance vector models that came with the code were tested on our images to classify the colour. Furthermore, the images were normalised with the RgbHist and NormReinhard method. The normalised images were saved and used to evaluate the normalisation performance and finally to select the combination of colour normalisation techniques and the classifier model suitable for our domain.

### 3.5.2 Result and conclusions

Referring to Table A.1 shows that not all normalisation techniques work well on all images, and yet some may still work with a different colour classifier vector model. The output of the colour normalisation with a different vector model are shown in Table A.2 and A.3. Experiments also reveal that images with sizes greater than 7000 pixels will not be able to proceed with all of the normalisation algorithm.

The results on a test set of images are shown as Appendix A.1. The colour normalised images, as in Tables A.2 and A.2, are observed particularly at the contrast between the background and tissue, similarity of image saturation and the impact of image sizes to colour normalisation techniques. The observations are as below:

- Reinhard

Images with darker nuclei become brighter with white background, but images with more background become too saturated and therefore the backgrounds which are supposed to be white change to slightly dark pink compared to the rest of the other normalisation techniques. This is caused by unassigned pixel classes in images so it is not feasible for H&E virtual slides images. Furthermore, the mean and standard deviation for each channel in the whole image are not a good representation for this kind of image as the colours are directly related to stain-specific saturation to show different kinds of protein.

- VB Reinhard Hard

Normalised images seem pinkish as well but images produced generally have a similar appearance. This is caused by the poor segmentation between the background and the cytoplasm which is coloured pink in the images.

- VB Reinhard Weighted

All normalised images have a slightly pinkish background, but the colour classifier that performs better than the rest would be liver2 and no barretts.

- CDV Linear

Images produced with this normalisation technique have even colour distributions



and similar appearance. The backgrounds are not changed and the colour saturations are normalised. However, the Normbimodal method has failed to normalised images which are more than 5000 pixel wide or/and high as well. Therefore, we have to reselect our ground truth images for smaller regions, making sure that our region selection module uses this threshold value as well if colour normalisation is to be implemented.

- **CDV Multimodal**

Normalised images produced by this colour normalisation technique produce uneven output, where some output images are too bright and other are too saturated. Therefore, we do not implement this technique in our future work as we need images with a similar appearance for better colour deconvolution results.

Thus, CDV Linear colour normalisation technique by [71] gives better results on our images compared to the VB Reinhard (original, hard and weighted) as well as obviously the CDV Multimodel. We also found that image size plays a major role in the normalisation process with these relevance vector models as Magee et al. [71] uses  $1000 \times 1000$  pixels in 40X magnification in their experiments, but our images are varied in size and compressed to 20X. The relevance vector model colour classifier also helps to provide a better colour normalised image without compromising the image contrast and background.

To test the robustness of Normbimodal normalisation technique further, an additional 40 annotated images sized  $<5000 \times 5000$  pixels with an epithelial layer of grade 1 and 5 were used. This time, all images were processed but the normalised images for some colour classifiers simply went blank or reddish, as summarised in Table 3.6.

rvm classifier	image blank/black	image red
liver4a	all	-
liver3p1	13(3016*3880)	-
liver4p2	13(3016*3880), 40(2792*2436), 55(976*956), 56(2052*1820), 101(1484*1086), 262(3408*2124), 627(1592*1186)	350(1954*1166)
liver3p2	56(2052*1820)	-
liver4p3	-	350(1954*1166)

Table 3.6: Normalisation failure in additional images

However, it took around a week to normalise 40 images with the CDV Linear (HE, liver 1, liver2, liver4p1 and no barretts colour classifier model) with a significant number

of failure cases, as shown in Table 3.7. From the normalisation process, only normalisation images from two colour classifier models (liver 2 and no barretts) are acceptable for our images, but risk failure of normalising the colour for large images. Due to image size and time constraints, we proceed with the image analysis without colour normalisation and use grey level images to reduce the impact of colour variations.

rvn classifier	image id	image size
liver1	214	4840*4480
	500	4608*4656
liver2	-	-
liver4p1	500	4608*4656
HE	214	4840*4480
	500	4608*4656
no barretts	-	-

Table 3.7: Bimodal normalisation failures on ground truth images

## 3.6 Conclusions

This research relies on two types of ground truth image: whole virtual pathology slides and the annotated regions marked and graded by domain experts. Although we have many images available, we can only make use of 75 virtual slides with [60:15] ratio for [train:test] images and [90:24] for annotated regions. This is mainly because the ground truths for dysplasia with grade 3 and grade 4 are limited. A balanced number of data for each grade is selected to avoid over-fitting, as explained in Chapter 2.6.1.

General pre-processing techniques, as explained previously in Chapter 2.3 were carried out on our training dataset. Noise removal processes and initial experiments were carried out to find the best way to create regions of tissue from virtual slides. We have found out that creating tiles of regions sized 1800 pixels wide  $\times$  1800 pixels high at 40X magnification across a tissue provides better images without significant loss of texture information.

Finally, we have run a quick experiment on colour normalisation techniques for our images to decide whether we should consider colour or just greyscale images in further processes. The effects, as in Chapter 3.5, show that it makes matters better when the colour normalisation works, but sometimes it does not work and this detracts from the overall performance. Furthermore, huge data size and the time taken for normalisation has contributed to the decision to use greyscaled images for tissue texture analysis, which will be explained later in Chapter 4.

## Chapter 4

# Annotated Region Tissue Analysis

---

The next step after image preparation and region creation is to extract textural features from each region. Generally, the processes involved are illustrated in Figure 4.1 where the bottom layer contains regions that we have prepared, as explained in the previous chapter.

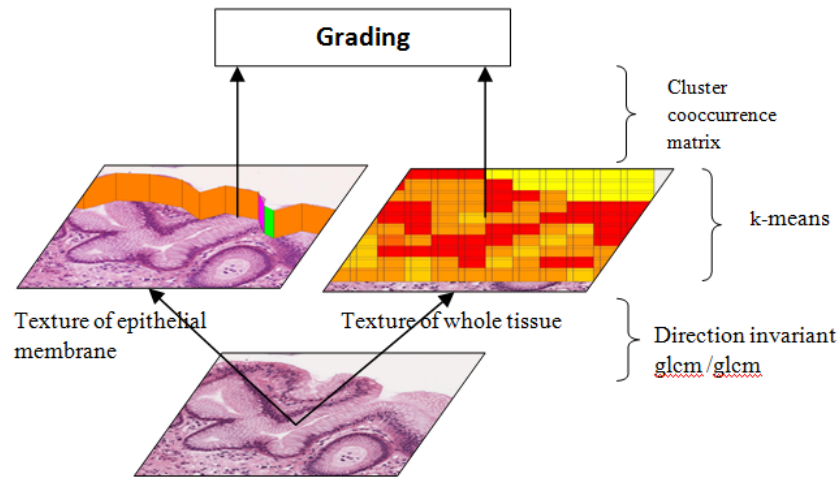


Figure 4.1: Region for texture extraction from annotated regions

This chapter contains the methods and experiments that we have carried out to extract textural features from the regions. Regions were processed with two parallel feature extraction processes: (i) at the epithelial membrane layer and (ii) the whole tissue in regions. The epithelial layer analysis focuses on extracting texture features of the surface

membrane. This includes experiments to investigate the significance of epithelial layer texture features in grading dysplasia, as pathologists always associate changes of cells in BO with the tissues on the surface or the epithelial layer (discussed in Chapter 2.1).

The whole tissue texture is analysed at 2 levels: the pixel-level and the cluster-level where the tissue image is transformed into another representation while preserving its spatial information. Patches were created across the region and GLCM features were extracted. Many brute-force experiments were carried out to find the best parameter settings as well as to select features. These features were used to clusters patches based on their texture similarities using non-supervised clustering. Then, the spatial features of the regions based on the clusters arrangements were used to grade regions, following the modified Vienna Classification of Dysplasia, as shown in Table 3.1.

Lastly, classification models using selected features from both epithelial layer and the whole tissue are investigated for implementation in the next process: whole virtual slides grading in Chapter 5.

## 4.1 Epithelial layer texture analysis

As [38, 75, 87] have reported that the epithelium layer is an important area to be examined by pathologists in characterizing dysplasia, we decided to include this in our research. Formulating the hypothesis:

$H_0$ : Texture features in the epithelial layer contain information to classify dysplasia and non-dysplasia.

$H_1$ : Texture features in the epithelial are not enough to classify dysplasia and non-dysplasia.

In order to prove the hypothesis, we selected 40 annotated regions containing the epithelial layer, where 20 annotated regions were non-dysplasia (G1 and G2), versus 20 annotated regions which were dysplastic (G3, G4, G5 and G6). To extract only the texture of the epithelial layer, the tissue surface (presumably the epithelial layer) has to be detected and differentiated against the muscularis mucosa.

However, there are two options to choose the baseline for the patches creation: (i) the nuclei lining or (ii) the region boundary detected with grey level thresholding and connected components (as discussed in Chapter 3 previously). Figure 4.2 shows the two different baselines accordingly. A mechanism to create patches along these baselines is used to extract texture at the pixel level. Each patch is clustered based on the texture feature similarities and the cluster arrangement in each region is used to grade the dysplasia.



(a) line following nuclei lining (b) line following tissue boudary

Figure 4.2: Two different baselines to choose as a reference point in creating patches.

#### 4.1.1 Reference point selection

To extract the epithelial tissue layer, two approaches were tested. The first uses the same method to detect the boundary as discussed in Chapter 3. Each annotated region is converted into a greyscale image and binarised. In this binary image, pixels with value ‘0’ are considered as background. Then, each pixel is scanned from the top left corner of the image to the bottom right corner, with a horizontal scanning movement. The first detected pixels with value ‘1’ will be the starting point for the boundary detection.

From the starting point, coordinates for each connected pixel in eight neighbourhood directions and adjacent to a ‘0’ pixel are recorded as a candidate boundary. Therefore, in each annotated region, we will have a set of lines which will end at the bounding box, but might start anywhere in the annotated regions. Thus, we exclude five pixels around the image to ensure that the bounding box is not included as the region boundary.

Furthermore, there are also be cases where part of the lines are detected twice but counted as a different set, as in Figure 4.3 a1. Therefore, only unique coordinates are retained, except for the start and end points. The start and end point for each line are compared, and if they share the same coordinate, the two lines are linked together as one. Then if this new line is the longest set of connected pixels, it will be accepted as the boundary, as shown in Figure 4.3 a2. The number of connected coordinates for each line, as well as the straight line distance between the start and end pixels, are also recorded. The algorithm to choose the reference line from the connected component is repeated for each annotated region.

The longest set of connected pixels is selected as a default for the reference line and marked with black. However, there are cases where the longest line might not be the epithelial layer, but the muscularis mucosa (as shown in Figure 4.3 b1) or the torn-off tissues at the epithelium (as in Figure 4.3 c1). These areas do not represent the correct behaviour of a dysplastic tissue, and become the longest detected pixels due to the complexity of the surface. In order to overcome this, we use the straight line distance between the start

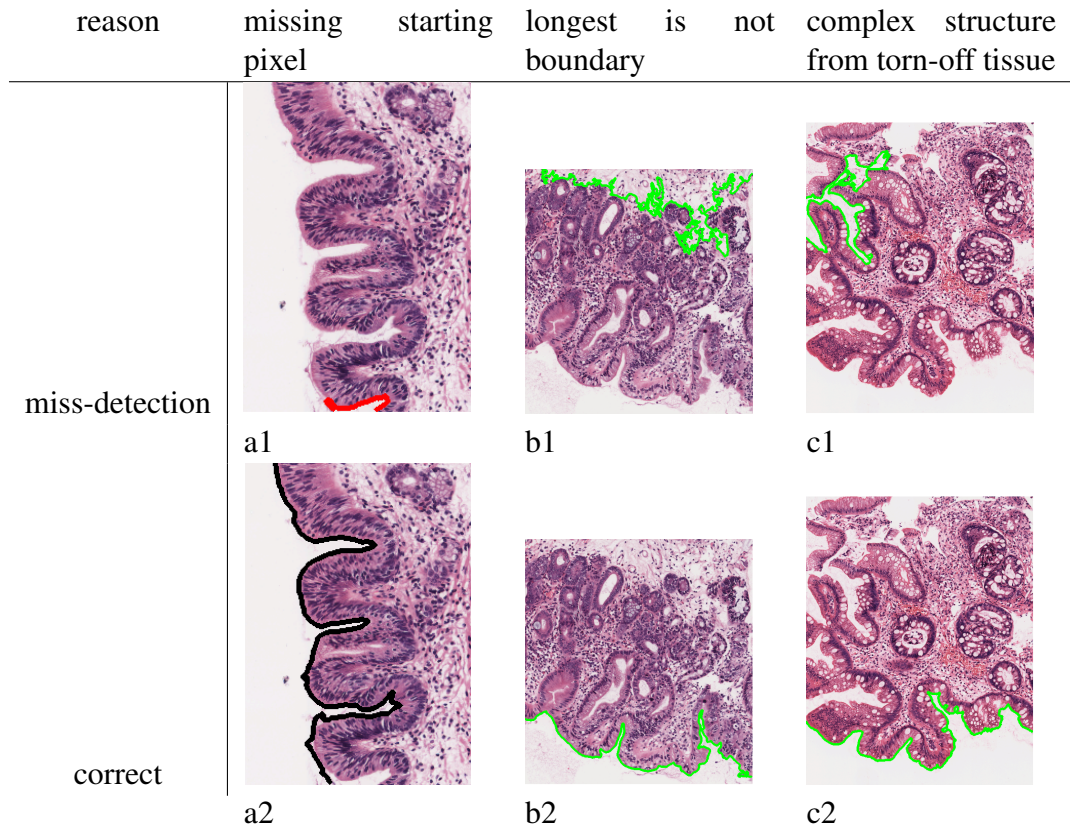


Figure 4.3: Sample of mis-detected boundary and the corrected output.

point and the end points for the second, third or fourth suggestion of tissue boundary for manual verification. Then the accepted boundary will be marked with green as in Figure 4.3 b2 and 4.3 c2.

Even though this method can detect tissue surfaces, it leaves out peaks and crypts in a very complex tissue structure or closed curved boundaries, risking some information loss.

The second reference point is the nuclei lining along the tissue boundary. Referring to the cytological changes discussed in Chapter 2, the location of nuclei in the columnar cell on the surface epithelium moves from the bottom middle to the top of the cells, but remains perpendicular to the basement membrane except for IMC. The thickness of the epithelial cells is highly affected by dysplastic conditions. Therefore, these reference points are believed to be better because we can still capture the epithelial layer, regardless of the cell's thickness.

To use the nuclei lining, we applied an optical density matrix in colour deconvolution for Hematoxylin and Eosin (H&E), using the optical density matrix suggested by Ruifrok in [103]. By applying this, a clear image with enhanced nuclei or cytoplasm is obtained,

making it easy to analyse the features. The image boundary is projected after the morphological operation is applied on the deconvolved image, as shown in Figure 4.4. It shows the original annotated regions stained in H&E (a), nuclei lining (b and c) and cytoplasm (d and e); before and after the morphological changes are applied.

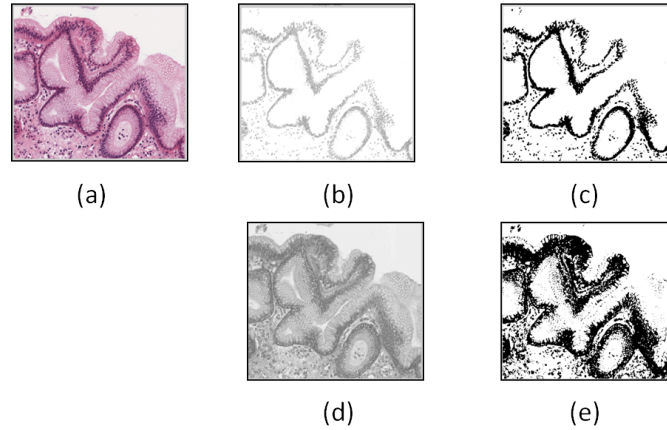


Figure 4.4: Colour deconvolution process to detect nuclei boundary as a reference point to create patches.

Morphological operations such as opening and closing are used to bridge every disconnected nucleus. Then skeleton and thinning methods are used to create a smooth line representing the nuclei lining in a tissue. This method is good at picking up crypts and peaks in complex boundary structures. However, the lines produced for highly stratified regions are too detailed, making it unsuitable for patch creation. On the other hand, no boundary lines are detected for normal and indefinite dysplasia tissue as the nuclei are not connected. It was therefore not considered not appropriate for an automated approach.

For experimental purposes, we use a hand-marked mask on the tissue by following the nuclei lining as closely as we can. Sample images of the original region, the detected epithelial layer based on tissue boundary (first method) and based on the nuclei lining (second method), as well as the hand-drawn mask (for comparison) are shown in Figure 4.5.

In Figure 4.5, we can see clearly that the first method managed to detect the correct tissue boundary, except for the last image. The failure is justified by the condition of the image, which is cropped on the tissue boundary but not affecting the nuclei lining. The second method fails to detect the boundary for the third and fourth annotated regions because the nuclei are interconnected. However, the detected boundary for the first image contains many false negative lines as the detected nuclei are large and crowded. However, an acceptable boundary was generated for the second image, thus it will be used in further



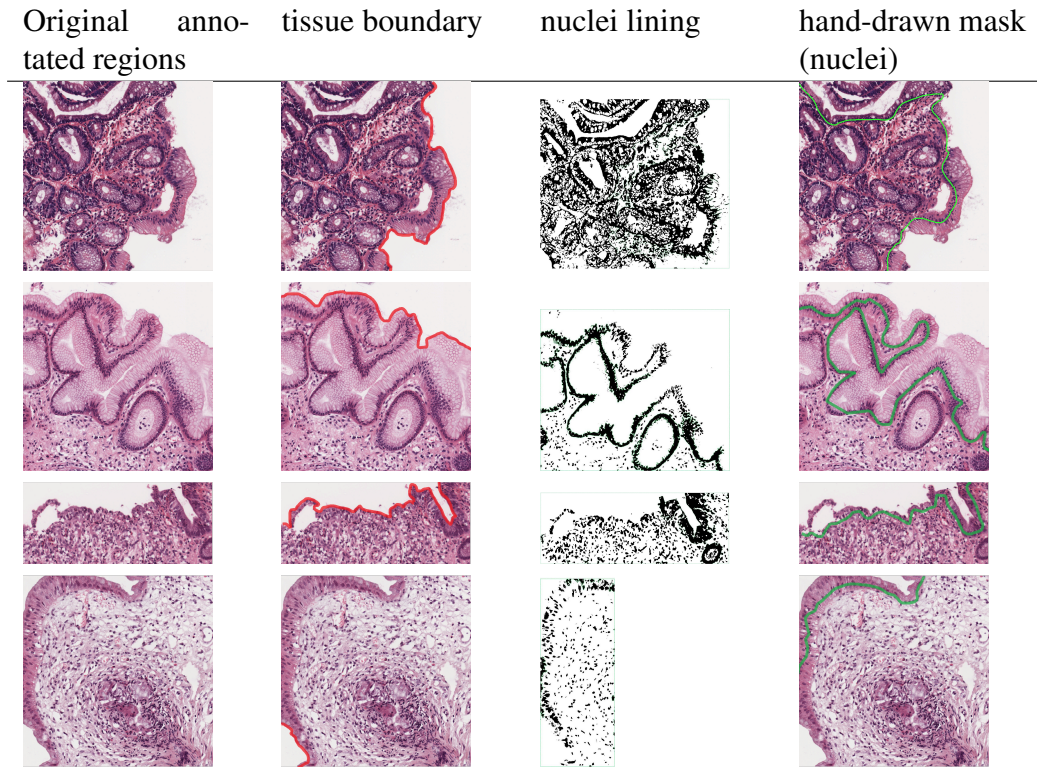


Figure 4.5: Sample annotated regions with three different boundaries.

processing.

#### 4.1.2 Patch creation

To extract features of the surface membrane, annotated regions are analysed at 40X, 20X and 10X magnifications. When tissue boundaries are curvy, so are the epithelial cells along the surface. Therefore, to investigate the effect of a cell's rotations on patch texture at pixel-level, rotated and un-rotated patches were created at each magnification as illustrated in Figure 4.6.

Based on the detected boundary as a reference line, coordinates for every  $j^{th}$  pixel are used as 'a sliding window' to create patches ( $P$ ) along the boundary. The slope between two successive  $j$  pixels ( $\delta$ ) was calculated and coordinates for generating patch  $P_1, P_2, \dots, P_z$  were projected, where  $z$  is the number of patches. Different patch height ( $R=50, 100, 150$  and  $200$  pixels) were used to extract the most useful sub-images for the next step.

Two directions of patches were evaluated, which are (i) perpendicular to the  $\delta$ , and (ii) perpendicular to the normal surface. For (i), patches are rotated based on the tissue direction to retain the spatial correlation of the tissue surface. This is carried out by



processing two successive points  $(j_n, j_{n+1})$  on the boundary coordinates as detected in Chapter 4.1.1.

As shown in Algorithm 1, different patch sizes ( $R$ ) were tested. The ‘ $R$ ’ is used to calculate new coordinates on both sides of the perpendicular line (of the normal surface or the  $\delta$ ),  $(R'_j, R'_{j+1}, R''_j, R''_{j+1})$  as in Figure 4.6. Then, the entropy value along points  $R$  to  $R'_j$  and  $R$  to  $R'_{j+1}$  are compared to ensure that patches are created on a tissue, not background.

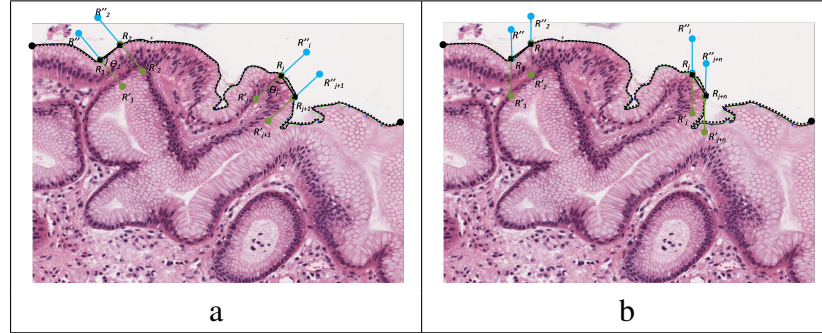


Figure 4.6: (a) shows the creation of rotated patches along the detected epithelial layer while (b) shows the un-rotated patches.  $R$  represent the height and  $\theta$  represents the angle for rotation.

The given algorithm will produce a line of patches along the reference line. The new patches will mostly contain the epithelial layer and are then ready for texture feature extraction processes. Samples of rotated and un-rotated patches with several  $R$  values are shown in Figure 4.7.

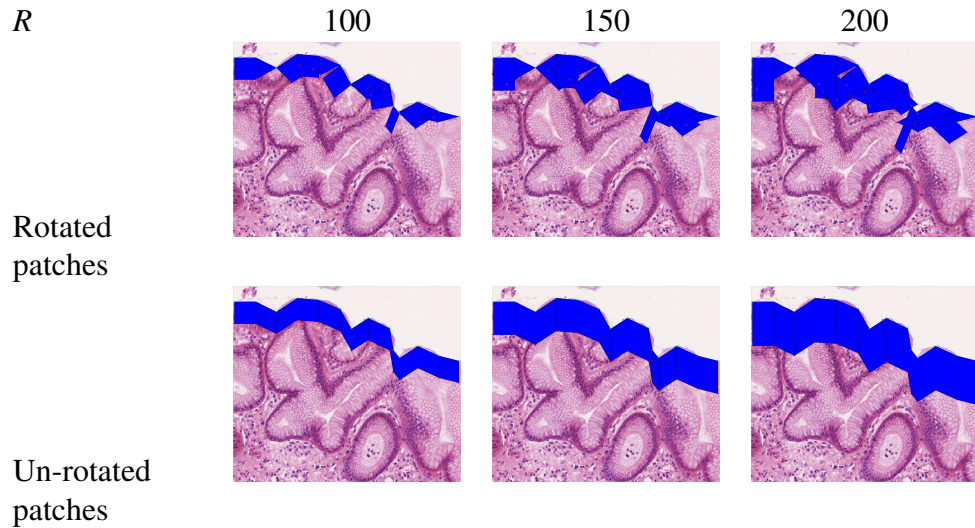


Figure 4.7: Figure showing the rotated and un-rotated patches with different  $R$ , created with the algorithm shown.

**Algorithm 1:** Patch creation along tissue boundary.**Data:**  $i=100$ ,  $R=[100, 150, 200]$ ,  $num=coordinate$ **Result:** Patches along reference line**forall the annotated regions do**     $j \leftarrow 1, r \leftarrow 1$     **while**  $j \leftarrow 1 \text{tolenght}(num)$  **do**        *coordinate selection*         $repm(r, 1) \leftarrow num(j, 1), repm(r, 2) \leftarrow num(j, 2)$          $r \leftarrow r + 1, j \leftarrow r * i$     **end**    **for**  $j \leftarrow 2 \text{tolenght}(repm)$  **do**         $h \leftarrow j + 1$          $R_{j(x)} \leftarrow repm(j, 1), R_{j(y)} \leftarrow repm(j, 2)$          $R_{j+1(x)} \leftarrow repm(h, 1), R_{j+1(y)} \leftarrow repm(h, 2)$         *for unrotated patches and rotated patches with slope ( $\delta$ ) = 0;*         $R'_j \leftarrow [R_{1x}, R_{1y} - R], R'_{j+1} \leftarrow [R_{2x}, R_{1y} - R]$          $R''_j \leftarrow [R_{1x}, R_{1y} + R], R''_{j+1} \leftarrow [R_{2x}, R_{1y} + R]$     **end**    *for rotated patches only, calculate the  $\delta$*      $\delta \leftarrow \text{slope between } (R_j, R_{j+1})$     **if**  $\delta = \infty$  **then**         $R'_j \leftarrow [R_{1x} - R, R_{1y}], R'_{j+1} \leftarrow [R_{2x} - R, R_{1y}]$          $R''_j \leftarrow [R_{1x} + R, R_{1y}], R''_{j+1} \leftarrow [R_{2x} + R, R_{1y}]$     **else if**  $\infty < \delta > 0$  **then**        *formulate transformation coordinate*         $pslope = \text{real}(-1/\delta)$          $theta = \text{real}(\text{atanh}(pslope))$          $yslide = \sin(\delta) * R$          $xslide = \cos(\delta) * R$          $R'_j \leftarrow [R_{jx} - xslide, R_{jy} - yslide]$          $R'_{j+1} \leftarrow [R_{(j+1)x} - xslide, R_{(j+1)y} - yslide]$          $R''_j \leftarrow [R_{jx} + xslide, R_{jy} + yslide]$          $R''_{j+1} \leftarrow [R_{(j+1)x} + xslide, R_{(j+1)y} + yslide]$     *calculate entropy*    **if**  $\text{entropy}(R'_j, R'_{j+1}) < \text{entropy}(R''_j, R''_{j+1})$  **then**         $patch \leftarrow (R_j, R_{j+1}, R'_j, R'_{j+1})$     **else**         $patch \leftarrow (R_1, R_2, R''_j, R''_{j+1})$     **end****end**

### 4.1.3 Feature extraction

For feature extraction, these patches are then converted into greyscaled images for pixel-level texture feature extraction processes. Four GLCM features namely contrast, correlation, energy and homogeneity in four directions (as explained in Chapter 2.4.1) were calculated to represent the patches texture.

There are many parameters to fine-tune in order to get the optimum features set representing our image. These include the patch height ( $R$ ), patch width ( $pz$ ), zoom level, number of cluster to use ( $k$ ), feature set to use and feature directions as well as offsets or neighbouring distance for GLCM ( $n$ ).

From the output of Algorithm 1, the first observation to choose the patch height ( $R$ ) is carried out. The patches should ideally cover the whole epithelial layer and not include the lamina region. From the output images (in Table B.1 and Table B.2 in appendix B),  $R = 150$  is selected as it produced the closest desired output, which is the epithelial layer only.

Next, the GLCM texture features (correlation, energy, homogeneity and contrast) were extracted from each patch. These patches are clustered based on the texture similarity. Using k-means clustering, we group these patches into several clusters automatically. The unsupervised clustering used Squared Euclidean Distance with several values of ( $k = 5, 7$ ) for evaluation purposes. Some samples of the clustered patches on annotated regions in 10X magnification with different  $k$  and  $R$  are shown in collection of annotated regions figured with rotated patches; Table B.1 and un-rotated patches; Table B.2.

The GLCM features of these patches were used to grade annotated regions into G1 or G5 with  $pz = 100$  and 150. In this experiment, 14 images for dysplasia G1 and another 14 for dysplasia G6 were used. As we can see from Figure 4.8, the result from  $pz = 150$  is more consistent with three GLCM features giving  $>50\%$  AP, regardless of the number of clusters,  $k$ .

The effect of zoom-level and neighbouring distance ( $n$ ) were further investigated using the same features as above. Therefore, patches were created on 10X and 20X magnification level, and the  $n$  tested were [2,4,6,8,10]. The results are presented in Table 4.1 showing that grading performance for  $k = 5$  is more reliable as the AP across a different number of  $n$  does not vary much, either at the same zoom level, or different zoom level. If  $k = 7$  is used, the AP performance fluctuates across different values of  $n$ , as well as a big classification difference between  $zoom = 10X$  and 20X. Based on the AP achieved and the mean squared errors ( $MSE$ ), zoom 10X is selected with  $n = 4$ .

Then, in order to test if patches provide direction invariant features, 17 patches with  $pz=100*100$  were selected and rotated to  $90^\circ, 180^\circ$  and  $270^\circ$ , producing 108 patches.

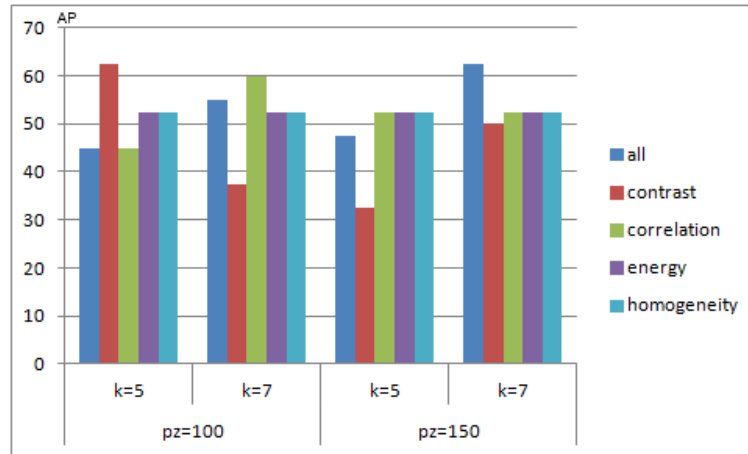


Figure 4.8: Grading annotated regions into G1 and G5 using different set of GLCM features,  $k$  and  $pz$ .

GLCM features in four directions were extracted from all these patches with  $n=10$ , producing 160 features (setA). SetB is the average of the GLCM features over four directions, producing 40 features, SetC is GLCM features only on the major directions (40 features) and last is SetD, which is the GLCM features over each direction (16 features).

Ten fold cross validation was used to see the similarity of these patches using DT and RF, and compared to a 70% split training-text to see the reliability and robustness of the result. As shown in Figure 4.9, features setB and setC are more rotational invariant as the texture features are averaged out over all four directions, compared with setC which uses only the major direction.

The DT used is the C4.5 algorithm (also known as J48) with confidence factor of 0.25 and the minimum number of objects in each leaf is two. The RF built up to 100 trees with a maximum depth of ten.

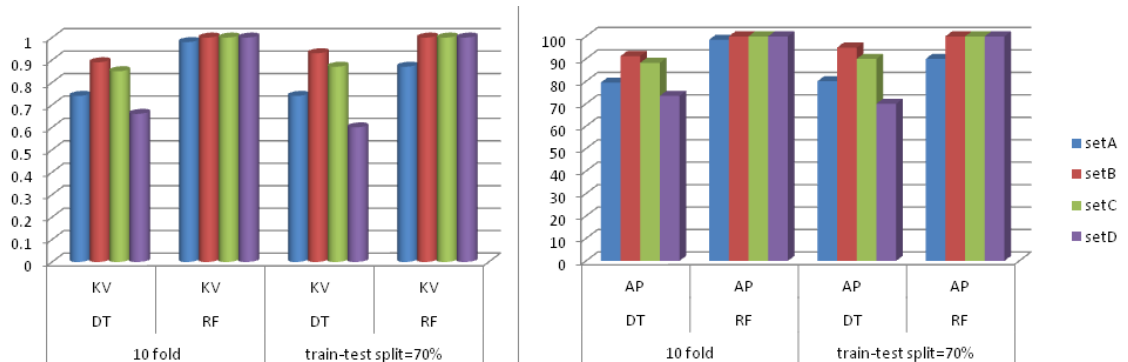


Figure 4.9: Graphs showing the AP and KV achieved with different sets of features, to test the effect of rotation for patches.

		$k=5$		$k=7$	
$n \backslash$ zoom		10	20	10	20
10	AP	75.0	71.5	85.7	67.9
	MSE	6.25	7.14	3.5	8.03
8	AP	78.6	78.6	82.1	67.9
	MSE	4.65	5.23	4.4	8.03
6	AP	78.6	78.6	60.7	85.7
	MSE	4.94	4.36	9.8	3.5
4	AP	82.1	78.6	64.3	57.1
	MSE	4.47	5.36	8.9	10.7
2	AP	78.6	71.4	71.4	67.9
	MSE	5.35	7.14	7.14	8.03

Table 4.1: Validation result on values of  $n, k$  and zoom level to grade annotated regions into G1 or G6 using texture of tissue boundary.

Three textural features extracted from rotated and un-rotated patches along the epithelial layer from both reference lines were evaluated to determine which one gives more information. The first texture is from the co-occurrence of clusters in one direction ( $0^\circ$  with two neighbouring patches) along the reference line. The second is the cluster frequency and the third is LBP features. The LBP features were included, to see if rotated and un-rotated patches have a similar performance with direction invariant texture features. All these features are used to grade annotated regions into dysplasia or non-dysplasia with 20 images for each class. The results are shown in Table 4.2.

Reference line	Input features		SVM	DT	RF
tissue boundary	cluster co-occurrence	rotated patches	75.0%	75.0%	75.0%
		unrotated patches	82.1%	75.0%	75.5%
	Bags of words	rotated patches	55.0%	60.0%	62.5%
		unrotated patches	72.5%	67.5%	80.5%
	local binary pattern	unrotated patches	55.0%	55.0%	60.0%
nuclei lining	cluster co-occurrence	rotated patches	47.5%	50.0%	32.0%
		unrotated patches	52.5%	42.5%	62.5%
	Bags of words	rotated patches	47.5%	42.5%	62.5%
		unrotated patches	47.5%	47.5%	40.0%
	local binary pattern	unrotated patches	33.2%	37.0%	40.0%

Table 4.2: Average result with 10-fold cross validation for comparison of rotated and un-rotated patches with  $pz = 150$ .

Referring to Table 4.2, we can conclude that texture features extracted from the un-rotated patches along tissue boundary reference line give better classification result. The

result also showed that the texture features from cluster co-occurrence performed better than the frequency of clusters (bags-of-words) and the LBP. Cluster co-occurrence features represent the spatial arrangement between different kind of tissue texture within a tissue while frequency of cluster only considers which clusters dominate the tissue as a whole. LBP, on the other hand did not perform well as the features are too many compared to the patch size, thus overfitting might have occurred (refer to detailed explanation in Chapter 2.4.1 regarding bag-of-word and LBP, and Chapter 2.6.1 for overfitting).

To test which reference line is better for representing the epithelial layer, the features selected were further tested with train-test and k-fold cross validation techniques. Table 4.3 shows the classification results using the SVM on features obtained from the patches. The SVM uses Radial Basis Function as the kernel with  $\sigma = 1$  and smoothness function is set to  $\frac{1}{n}$ , where  $n$  is the number of features used (depending on which input feature was selected (refer to Table 4.2)). From Table 4.3 we can see that setting patches along the tissue boundary detected from the connected component approach gives 83.5% AP, a much better result compared to the nuclei lining which is only 65.0% AP from colour deconvolution.

Reference line	train-test (AP)	10-fold (AP)	8-fold (AP)	6-fold (AP)
Nuclei lining	65.0%	47.5%	45.0%	55.0%
Tissue boundary	83.5%	82.5%	82.0%	80.0%
Detail result 40	True positive 18	True negative 15	False positive 4	False negative 3

Table 4.3: Comparison of grading annotated regions into dysplasia and non-dysplasia between tissue texture along nuclei lining and along tissue boundary with  $k = 5$ .

The four false positive results were from mistakenly detected boundaries where the longest boundary is the lamina or torn-off tissue, rather than the epithelial layer itself. The same table shows the false negative annotated regions, where the regions are graded with a higher grade of dysplasia because of the complexity of the detected boundary is not the epithelial layer. The epithelial layer is not curvy any more as the nuclei are already large and had arranged themselves at the surface of the cells.

#### 4.1.4 Results and discussion

We have tested combinations of necessary parameter settings, in an attempt to come out with a model for epithelial layer texture extraction. The results shown in Table 4.4 are the

top AP that we can get out of all the possible parameter combinations. It summarises all values for parameters used, as well as the selected value based on our experiments.

Parameter	tested value	selected value
number of cluster $k$	3,4,5,6,7,8	5
patch height $R$	50,100,150,200	150
patch width $p_z$	50,100,150	100
Neighboring pixel(glcm) $n$	2,4,6,8,10	4
zoom	10X,20X	10X
reference line	nuclei lining, tissue boundary	tissue boundary
patches	rotated, unrotated	unrotated

Table 4.4: Parameters tested and selected values for epithelial layer analysis.

The parameters used for feature extraction are very important as they directly influence the grading output. Using the selected parameters, we managed to make use of the information on the tissue boundary, which is important in reviewing pathology slides [38, 75, 87]. This finding has contributed to the thesis as it solves the 'border effect' that has been an issue for some of the previous works [23, 32, 55, 60, 112].

We managed to get 82.5% AP to classify annotated regions into dysplasia or non-dysplasia. Therefore, we accept the  $H_o$  hypothesis that texture features on the epithelial layer can be used to classify dysplastic and non-dysplastic region.

## 4.2 Tissue texture analysis

BO and dysplasia cannot be determined only by a single cell, certain type of nuclei or by the existence of other local tissue structures as we have discussed in detail in Chapter 1.1 and Table 2.2 with all the clinical references. It is identified with relation to the surrounding tissue structure and therefore, a fine-grained texture or pixel based classification will fail. In order to capture the general textural features from the images, three main steps were carried out in every annotated image. The first step includes patch creation and clustering; to extract texture features of each patch and label them by clustering based on texture similarities. The second step is to generate a representation of the texture of the whole tissue using the patch labels and the final step is to classify dysplastic images using these generalised texture features. This is illustrated in Figure 4.10.

Tiles of partly overlapped patches sized  $50 \times 50$ ,  $100 \times 100$ ,  $150 \times 150$  and  $200 \times 200$  pixels were created across the whole annotated regions. GLCM matrix features in all four

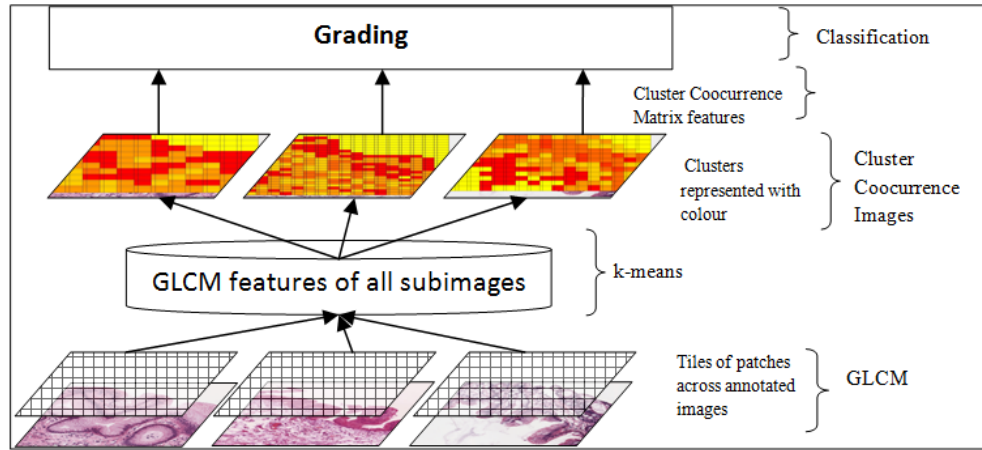


Figure 4.10: Texture features extraction steps from the lamina propria.

directions were calculated, producing 160 features for each patch. These features are used to cluster patches into an optimal number of clusters, enough to generalise patterns in the annotation images.

However, we found out from our experiments that many patches, especially from G5, have no pixel correlation value (NaN). Referring to our discussion at page 16 and 17 of this thesis, correlation is NaN when the images are homogeneous (all values are the same). Referring to Equation 2.2b, the standard deviation ( $\sigma$ ) of the two pixels is zero. Any number divided by zero will produce 'NaN' and that is the reason for the 'NaN' value in our correlation features here.

Therefore, we tested two conditions for this research. The first experiment (EXP1) is to cluster patches with all GLCM features (contrast, correlation, energy and homogeneity) with all no correlation values (NaN) changed to a very small positive correlation value (0.0001) to enable clustering. The second experiment (EXP2) is to cluster patches from image GLCM features without the correlation feature.

In order to get the optimum texture features, combinations of parameters for setting the zoom level, patch size, neighbouring size, number of clusters and variable texture features were fine-tuned. This includes varying the number of neighbours  $n$  between five and fifteen, and number of clusters  $k$  between three and eight, adding up to 800 experiments at each zoom level.

The best clustering achieved based on the distance between the cluster centroids was obtained when  $k=5$ ,  $n=10$  and  $p_z=100\text{pixel width} \times 100\text{pixel height}$  at 20X magnification. A very good clustering output was achieved when every patch was clustered to either C1, C2, C3, C4 or C5 based on the similarity of their GLCM features. Thus, we use the cluster centroids as our clustering model to cluster the test-set of annotated region dataset.



Referring to Figure 4.11, we can see that C1 contains patches with ‘picket-fencing’ texture, C2 contains texture where half is transparent tissue, and the other is nuclei. C3 contains patches with a coarse or high-grained looking texture, C4 contains images with rough-looking texture and C5 are blank patches which probably are the annotated regions’ background.

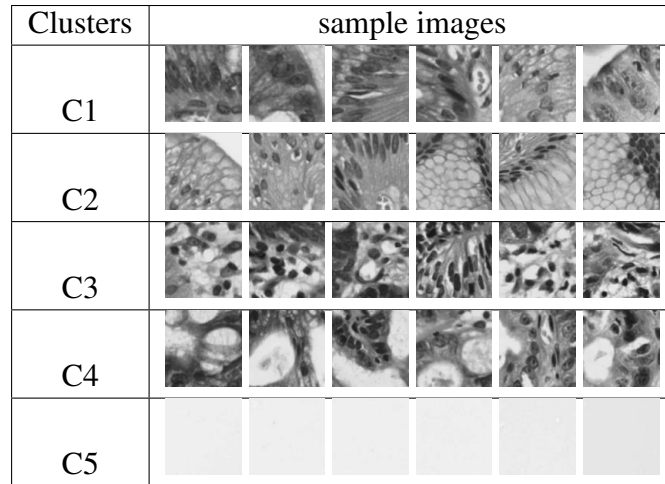


Figure 4.11: Sample of clustered patches with  $p_z=100*100$ , zoom=20X and  $n=10$  from EXP1.

Further experiments were carried out to grade dysplasia with a decision tree into dysplasia grades G1, G3 and G5 (refer to Table 3.1 for definitions) using frequency of clusters (also known as the bag-of-words method). Test results are shown in Table 4.5, showing that grading into three grades produced low classification accuracy (66.7% AP). The confusion matrix shows that frequency of clusters can be used to grade G5 as there is only one mis-grading case, but not for G1 or (especially) G3. This suggests that frequency of certain textures in tissue is not enough to differentiate between grades, and spatial information might be useful. Secondly, the grading could be improved with binary classification.

Features	best result	Confusion matrix			
			G1	G3	G5
Frequency of clusters	66.7% KV=0.17	G1	5	2	1
		G3	2	3	3
		G5	0	1	7

Table 4.5: Grading CCI with decision tree into G1, G3 and G5. Current agreement between pathologist is 0.24KV.

To investigate the effects of clusters on certain grades of dysplasia, a simple experi-

ments using binary classification was carried out using the frequency of clusters to grade dysplasia using an SVM. Based on the results in Table 4.6, C4 is an important cluster to differentiate between IMC and non-IMC as they give the highest accuracy compared to the other grades. The texture of cluster C3 is significant in differentiating between HGD and IMC while cluster C5 is between dysplasia and non-dysplasia. However, this is only a preliminary experiments. Detailed experiments with parameters will be discussed later in this chapter.

Grades	$k$ value	Average AP	Average KV	Cluster
HGD vs IMC	3,5,6,7	69.27	0.39	C3
IMC vs Non	3,5,6,7	81.06	0.37	C4
Dysplasia vs non dysplasia	3,5,6,7	64.20	0.24	C5

Table 4.6: Significant clusters of textures to classify certain grade.

Therefore, clusters are significant in grading dysplasia in annotated regions as it represents similar tissue texture. These clusters were then used to map the texture and spatial features between different types of tissue texture within the same annotated regions.

### 4.2.1 Cluster co-occurrence images

Cluster model as in Figure 4.11 was used to clustered together, patches with similar textures. Any new patches that are far from the existing centroids will be assigned to C6, the unknown cluster. Then, clusters are re-assigned to their original location in annotated images to create their Cluster Co-occurrence Images (CCI).

A CCI is generated by representing every patch with a pixel of the patch's clusters value. Therefore, the general texture and spatial pattern of each annotated image is preserved and the size is downscaled. Different from texon- based or bag-of-words approaches, the CCI is used to produce another level of textural and spatial features in the form of an image, rather than only a histogram of clusters. To enable visual examination of the relationship and the similarity of the generated CCI image with the original annotated regions, patches are assigned colour according to its cluster.

Figure 4.12 shows sample of CCI both from EXP1 and EXP2 with  $k=5$  and  $pz=100 \times 100$  at 20X magnification. Patches were created across the whole annotated regions (as the original images in Figure 4.12) and GLCM features for EXP1 and EXP2 were extracted for each one of them. These features were combined with other patches from all other images, and used to cluster them into five clusters of patches. Patches in each cluster were represented by the same colour, thus different patches from different clusters will

have different colour. These coloured patches were re-assigned to their original location at the respective annotated image, thus the CCI images as in Figure 4.12 EXP1 CCI and EXP2 CCI.

These two experiments produced a significantly different image. For example, EXP2 CCI of the first image in Figure 4.12 shows more colour variation of tissue clusters in the tissue area, and better detection of crypts compared to EXP1 CCI. More sample of CCI generated from different value of  $k$ , or at different zoom level are shown as Appendix C, Figure C.1 and Table C.1.

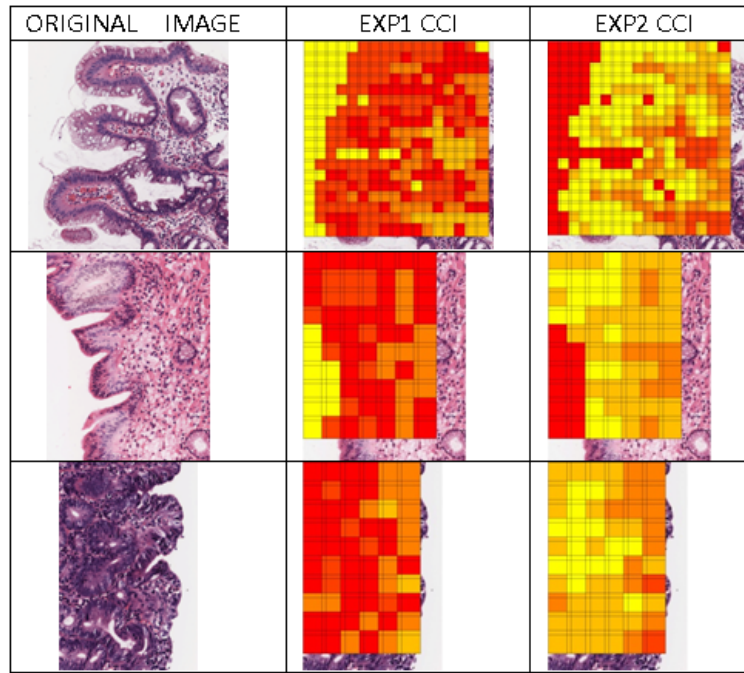


Figure 4.12: CCI generated from EXP1 and EXP2. We can see the difference between images produced from both experiments.

The CCIs should represent the original image structure, and the clusters are enough to differentiate each patch texture. From Figure 4.12, we can see a direct translation of the texture and spatial relationship within the CCI as in the annotated region itself. In addition, the CCIs are much smaller than the annotated regions as each pixel actually represent the GLCM texture features of a patch; which is actually  $100 \times 100$  pixels.

#### 4.2.2 The Cluster Co-occurrence Matrices

In order to investigate the possibility of having patterns of cluster arrangement in any particular grade of the same regions, the relationship between clusters was extracted. Thus, a Cluster-coded Co-occurrence Matrix (CCM) was extracted from each CCI. As GLCM

counts the co-occurrence of pixels value and the relationship between pixels, the CCM counts the co-occurrence of cluster types, as well as the relationship between them. The set of seven CCM features as listed in Table 4.7 were extracted from EXP1 and EXP2.

Set name	Explanation
A1	full <i>ccm</i> features (contrast, correlation, energy and homogeneity)
A2	average of <i>CCM</i> over all four direction
A3	average of <i>CCM</i> over each direction
A4	<i>CCM</i> features without correlation
A5	average of A4 over all four direction
A6	average of A4 over each direction
A7	frequency of clusters in each <i>CCI</i>

Table 4.7: CCM textural features derived from CCI.

However, before the CCM features were extracted, the largest group of neighbouring C4 were deleted from the co-occurrence matrix, as to remove the image background. C6 which is the unknown clusters are also deleted.

Initial testing on the value of  $n$  and zoom level is carried out with a SVM, to grade annotated regions into G1 and G6. The result is shown in Table 4.8, where the highest APs achieved came from zoom 20X with  $n=4$ . Value of  $k=5$  is following the clustering model, but with additional C6 as unknown cluster.

$n \backslash$ zoom	10	20	40
10	67.4	62.8	68.6
8	77.9	60.5	60.5
6	84.9	76.7	60.0
4	84.9	97.9	69.8
2	77.9	74.4	66.3

Table 4.8: AP for testing results on values of  $n$ ,  $k$  and zoom level to grade annotated regions into G1 or G6 using CCM features.

Then, using a brute-force approach, experiments to test the grading result using all possible combinations of parameter settings was carried out. One reason is to see the CCM features' ability to grade dysplasia into several classes using DT and RF, and the result is shown in Table 4.9.

As shown in Table 4.9, classifying regions into five different grades using level of texture features alone did not provide satisfactory results. However, the result improves when the number of classes is reduced to three and two classes. The results also depend

on the grades of dysplasia used. This is in line with interobserver variations research in [27, 91, 97].

No of classes	Cases	Classifier				Literature review
		<i>DT</i>		<i>RF</i>		
		<i>AP</i>	<i>KV</i>	<i>AP</i>	<i>KV</i>	
5	G1 vs G2 vs G4 vs G5 vs G6	31.13	0.14	28.2	0.1	0.25-0.27 [56]
3	G1+G2 vs G4 vs G5+G6	47.41	0.21	44.44	0.17	0.24 [119]
2	G5 vs G6	71.47	0.42	74.13	0.47	0.42 [119]
2	G6 vs G1+G2+G4+G5	79.53	0.41	81.47	0.37	*
2	G1+G2 vs G4+G5+G6	72.40	0.38	67.00	0.31	0.33 [56]
2	G4 vs G5	62.05	0.24	63.27	0.26	*
2	G4 vs G1+G2	72.4	0.38	63.27	0.24	<50%
2	G4 vs G1+G2+G5+G6	69.73	0.09	72.67	0.03	72% [119]
2	G5 vs G1+G2+G4+G6	78.60	0.00	75.27	0.14	*

Table 4.9: Comparison of classification between Decision Tree and Random Forest achieved at 20X. (\* not found in literature)

The AP obtained to classify five classes on average is only 31.1%. When reduced to three classes, the result increased to 47.4% on average. Then, the number of classes is further reduced to two, the AP achieved has dramatically increased. This is especially between two grades which are just next to each other.

Poor results were obtained in distinguishing texture between LGD and non-LGD, as well as between HGD and non-HGD. This is because both LGD and HGD are in the middle of the transition process from non- dysplasia to becoming IMC. The texture of these two grades is believed to have a combination of classes from either side of them in the transition process, thus differentiating them alone from the whole grading is not quite fair.

The results achieved have proved that tissue changes across few grades of dysplasia in BO can be measured using only texture features. This is because the KV achieved in diagnosing certain grades is equal or higher than the KV achieved by the pathologists themselves.

The CCM standard features and the frequency of clusters were used to grade dysplasia in each CCI with SVM with variety of  $k$  values to be tested. The classification results for the whole regions CCM texture features and the cluster frequencies are shown in Table 4.10. Both features were classified by the SVM, and for comparison purposes, a BDT and RF classifier were also used. The best classification result is obtained from CCM features

on SVM using our centroids as a clustering model, but with an additional C6: unknown cluster.

Features	correlation	$k$	SVM	BDT	RF
CCM	excluded	7	75.0%	75.0%	75.0%
		6	82.5%	75.0%	77.5%
		5	70.0%	67.5%	80.0%
	Nan=0.0001	7	55.0%	60.0%	62.5%
		6	72.5%	67.5%	80.0%
		5	67.5%	57.5%	72.5%
cluster frequency	excluded	7	55.0%	75.0%	68.0%
		6	55.0%	72.5%	75.0%
		5	55.0%	75.0%	77.5%
	Nan=0.0001	7	52.5%	77.5%	85.0%
		6	50.0%	75.0%	70.0%
		5	55.0%	77.5%	75.0%

Table 4.10: Grading AP with  $pz=100*100$  on different values of  $k$ .

### 4.2.3 Result

Based on many experiments that have been carried out, we have managed to use a clustering model to cluster all our patches based on texture similarities. The parameter values tested, are shown in Table 4.11.

level	parameter	tested values	selected value
pixel-level	k	5,6,7	5
	pz	50, 100, 150, 200	100
	n	5 to 15	10
	zoom	10X, 20X	20X
patch-level	n	10, 8, 6, 4, 2	4

Table 4.11: Tested and selected value for parameter setting for whole annotated region texture analysis

All seven sets of CCM features, as detailed in Table 4.7 from EXP1 and EXP2 were used to train the BDT, specifically to grade each image which is G1 or not-G1(Tree-G1), G3 or not-G3 (Tree-G3) and G5 or not-G5 (Tree-G5). Ten fold cross validation was used to validate the robustness of the features extracted and a validation set consisting of 24 unseen annotated images is used to test the performance of the decision tree models.

The grading results were compared to the ground truth; which is the grade given by the consultant pathologist when they create the annotated regions. Two values were used

to measure classification performance; the AP and the KV. KV is used by pathologists to assess the degree of interobserver variation in pathology; where  $<0.21$ ,  $0.21-0.40$ ,  $0.41-0.60$ ,  $0.61-0.80$ , and  $>0.80$  are commonly accepted interpretations of poor, fair, moderate, good and very good agreement respectively.

Patch features	EXP1			EXP2		
Grade	G1	G3	G5	G1	G3	G5
Features	A1	A4	A1	A7	A3	A6
AP	75.0	75.0 [72.0]	81.3	87.5	75.0 [72.0]	75.0
KV	0.5 [0.33]	0.5	0.63 [0.6]	0.75 [0.33]	0.5	0.5 [0.6]
confusion matrix	7 1 3 5	6 2 2 6	6 2 1 7	8 0 2 6	4 4 0 8	7 1 3 5
precision	0.88	0.75	0.75	1.00	0.50	0.88
recall	0.7	0.75	0.86	0.80	1.0	0.70

Table 4.12: Test result for CCM features selection from EXP1 and EXP2.

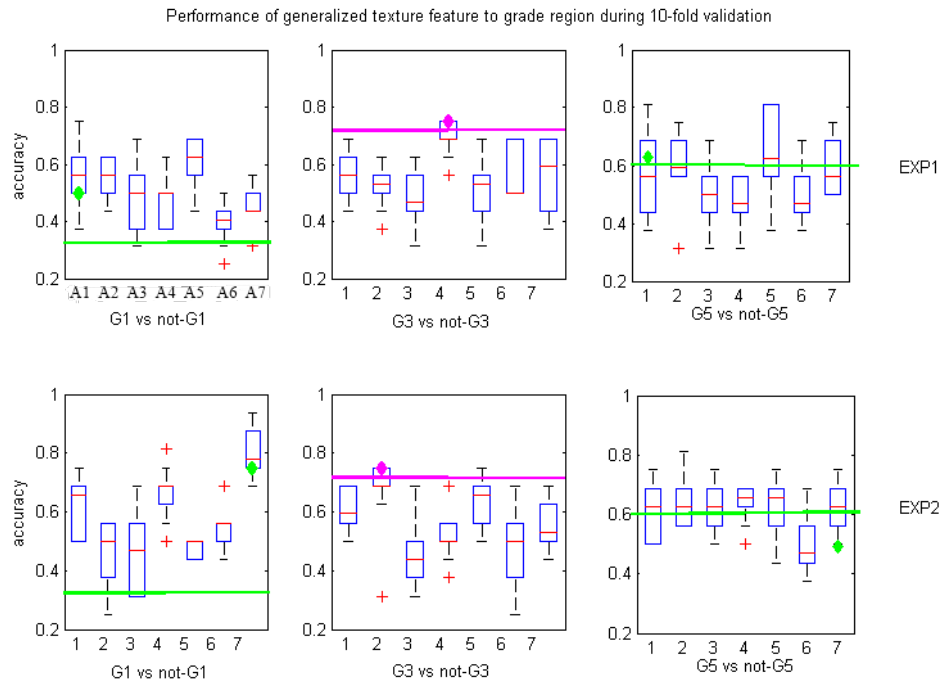


Figure 4.13: Comparison of binary tree grading result between different sets of textures in EXP1 and EXP2.

The best grading result achieved from our test data using EXP1 and EXP2 with BDT is shown in Table 4.12, and the reliability and robustness of our CCM texture features to grade dysplasia are shown in Figure 4.13. In this figure, the highest KV achieved from all

trees is represented by the green diamond marker while pathologist agreement KV were shown as the green line. The magenta diamond represents the highest AP achieved, to compare with current AP achievement (the magenta line).

Feature selection as shown in Figure 4.13 and Table 4.12, shows that grading G1 and not-G1 is the easiest case where the pool of grading results from all feature sets, both from EXP1 and EXP2 outperformed the existing agreement score where the highest KV achieved is 0.75 compared to 0.27 as reported in [56].

Grading G3 however, again demonstrated that this grade has the fuzziest texture, being in between BO with no dysplasia (G1) and BO with severe dysplasia (G5). Figure 4.13 shows that all set of features from both EXP1 and EXP2 has low performance in grading G3, compared to the existing agreement by pathologists. However, one of the trees with the highest performance shows that CCM feature set A4 of EXP1 and CCM feature set A3 of EXP2 achieved 75% AP, slightly better compared to 72% AP achieved by the pathologists.

The pixel level correlation value appears to be an indicator for G5. This can be concluded by comparing grading results from EXP1 and EXP2. EXP1 contains correlation features at pixel level whereas EXP2 does not. Feature set A1-A3 in both experiments contains the correlation feature between clusters while feature set A4-A6 does not. Generally results from features of EXP2 show consistently good accuracy compared to those of EXP1, showing that correlation feature at pixel-level significantly reduces the grading performance. However, the highest agreement scored from EXP2 is low compared to EXP1.

To evaluate the features sets, the best features selected was used to build and train a decision tree. A validation data set consisting of 24 unseen annotated images is used to validate the performance of the decision tree models, and the features respectively.

For candidate features from EXP1, feature A1 to grade G1 shows an inversely proportionate result with the number of test data. On the other hand, grading results for both G3 and G5 suggest features A4, A1 and A2 as grading features. The decision tree built with each failed to achieve any agreement with the validation data set.

EXP2 on the other hand gives a consistently good grading result both in validation and test data, thus EXP2 features will be used for the rest of this research. The feature validation test suggests that A7 may be used to generalise texture features for G1 and G5 while A2 may be used for G3.

Table 4.13 shows both validation and test results from Tree-G1, Tree-G3 and Tree-G5 with the selected features. Tree-G1 is the BDT to grade G1 and non-G1, Tree-G3 is the BDT for G3 and non-G3 while Tree-G5 is for grade G5 and non-G5 respectively.



These BDTs were selected as a grading model and shown as Figure D.1, D.2 and D.3 in Appendix D.

Evaluation criteria	Tree-G1 (A7)		Tree-G3 (A2)		Tree-G5 (A7)	
	validation	10-fold	validation	10-fold	validation	10-fold
AP	79.2	75.0	79.2	79.0	75.0	70.8
KV	0.54	0.49	0.58	0.56	0.49	0.39
recall	0.67	0.8	0.62	0.62	0.58	0.55
precision	0.75	0.88	1	0.88	0.88	0.75

Table 4.13: Grading performance of the selected CCM features to grade dysplasia in regions into G1, G3 or G5.

Tree-G1 (as shown as Figure D.1), were using feature set A1 of EXP2, ()that is without correlation feature at pixel-level). The root node of Tree-G1 model starts by comparing the number of cluster C5 in each CCI against the threshold value. As we have removed the biggest group of connected clusters C5 from CCI, the remaining might be a blank area in the tissue itself. Then the number of clusters C1 are used, three times in the decision tree. So, cluster C5 and cluster C1 are an important texture for G1. The number of cluster C3 and C2 only appear in one rule each.

BDT model Tree-G3 built with feature set A4 of EXP1 (as shown as Figure D.2), are CCM features with correlation at both pixel and cluster level. However, the correlation in cluster level has been pruned, leaving only contrast and homogeneity value, but at different directions. The root starts with the contrast feature at  $135^\circ$  to decide if a CCI is not-G3. Then the decisions rules use contrast again, at  $0^\circ$  and  $135^\circ$ ; and homogeneity at  $45^\circ$  and  $135^\circ$  direction to further decide between G3 and not-G3. The repeating usage of contrast value between texture types at  $135^\circ$  shows that it is important to mark a G3 in CCI.

The selected model for tree-G5 (as shown as Figure D.3) uses feature set A1 from EXP1, which means the CCM features extracted from for this model came from CCIs with the enforced correlation (NaN=0.0001) at pixel level. The correlations between clusters at region level are also included. As we can see from the tree, all CCM features (at different directions) are included in the decisions rules, except for the energy. The root starts with correlation at  $45^\circ$  and the tree is balanced on the left and right side, showing that there is no one straightforward rule to decide a G5.

To evaluate the BDT models built for each grade, the same feature is tested for grading using a SVM and RF with WEKA [39]. The best setting has been applied is with 100 trees with 10 initial seeds for RF. As for SVM, the best setting was using Sequential Minimal Optimization for separating hyperplanes and the kernel function is Gaussian Radial Basis

with  $\sigma = 1$ . The smoothness value is  $\frac{1}{n}$ , where  $n$  is the number of features used (depending on which feature set was selected). The evaluation was carried out using the 10-fold cross validation and the grading result from the test data is shown in Table 4.14.

Machine learning	Grade	EXP	AP(%)	KV	precision	recall
SVM	G1	EXP2 A7	50.0	0.0	0.375	0.500
	G3	EXP1 A4	50.0	0.00	0.875	0.500
	G5	EXP1 A1	62.5	0.25	0.750	0.600
RF	G1	EXP2 A7	81.0	0.63	0.625	1.000
	G3	EXP1 A4	68.8	0.38	0.750	0.667
	G5	EXP1 A1	75.0	0.50	0.625	0.833
BDT	G1	EXP2 A7	87.5	0.75	1.000	0.8000
	G3	EXP1 A4	75.0	0.50	0.750	0.750
	G5	EXP1 A1	81.25	0.63	0.750	0.857

Table 4.14: The grading performance of CCM features on a validation data with SVM, RF and BDT

From these three machine learning methods, it is obvious that SVM failed to differentiate G1 and G3. However, the result from RF closely follows the results from BDT where it consistently gives a good grading result for G1, G3 and G5 with lower recall and precision values. This is expected as the nature of both learning algorithms is very similar. It also means that cues for grading regions of BO virtual slides can be easily understood in human language as set of rules.

#### 4.2.4 Discussion

We have managed to grade regions of dysplasia into G1, G3 and G5 at average of 77.8 AP with 0.54 KV interobserver agreement with expert GI pathologists. This is significantly better, compared to the latest report on clinical assessment of dysplasia in [56] which is 0.27 KV by six expert GI pathologists.

The results support our hypothesis; that grading dysplasia requires the ability to generalise the distributions of different tissue textures in a region. This is demonstrated by the similarity of the CCI to its original annotated image. The produced CCI images have replaced all texture features on a pixel-based level into cluster-levels, but the spatial dependency remains.

Another interesting pattern that we can understand from the CCI is regarding the correlation features for G5. Initially, correlation features at pixel-level for G5 mainly produced ‘NaN’ value which means that there is no correlation between pixels (as explained

earlier in Chapter 4.2), thus we explore more with EXP1 and EXP2 to see if correlation feature can be used to identify G5. EXP1 contains the enforced small positive correlation at the pixel-level where all 'Nan' values were replaced with 0.0001 as explained in 4.2.

The best feature set are the full CCM features which contains the contrast, correlation, energy and homogeneity between clusters. Thus, it has proven that correlation at pixel level can be a good indicator for grade G5 as six out of seven CCM feature sets from EXP2 consistently return good classification performance in both test and validation data. Grading performance from features of EXP1 generally perform poorly with average classification is below the current agreement. Therefore, we can conclude that correlation features between pixels for G5 is not a significant feature (or is very subtle), but became stronger between clusters (type of tissue textures).

### 4.3 Spatial feature analysis

In an attempt to understand if clusters of texture have a significant relationship with the surface membrane, the distance of clusters from the epithelial layer is calculated.

Based on the frequency histogram, three bins of distances were identified; surface (tissue closest to tissue boundary), middle (tissue in lamina propria) and deep (furthest from epithelial tissue, deeper lamina propria and may include muscularis mucosa). These layers were created according to the distance of patches with the nearest tissues boundary where the surface is  $<200$  pixels, middle is between 200 pixels and 500 pixels and deep for patches which are  $>500$  pixels away from the epithelial layer.

As we have five clusters in CCI, the total number of these features is 15, where each cluster is divided into three groups, near (n), middle(m) and deep (d). Thus, we will have  $C1_n, C1_m, C1_d, \dots, C5_n, C5_m, C5_d$ . In addition, the frequency of each cluster in these three layers is calculated and used as another feature to support dysplasia grading. DT is used and the result is compared to RF and SVM.

#### 4.3.1 Results and conclusions

The above mentioned features are used to grade dysplasia using CCI, to test if it contains any pattern of dysplastic tissue. The average result from ten-fold validation, to grade CCI into G1, G3 and G5 is 49 AP with 0.2 KV; only slightly better than the CCM features (47.41% with 0.21 KV as in Table 4.9). Therefore, we apply BDT approach on these features, to grade each CCI into G1 or not-G1, G3 or not-G3 and G5 or not-G5. The best result of the ten-fold validation are as in Table 4.15.

Machine learning	Grade	EXP	AP(%)	KV	precision	recall
SVM	G1	EXP2	56.3	0.12	0.5	0.63
	G3	EXP1	50.0	-0.16	0.88	0.58
	G5	EXP1	37.5	-0.33	0.13	0.25
RF	G1	EXP2	56.3	0.12	0.50	0.57
	G3	EXP1	56.3	0.12	0.63	0.56
	G5	EXP1	56.3	0.03	0.5	0.57
BDT	G1	EXP2	75.0	0.50	0.75	0.75
	G3	EXP1	68.8	0.37	0.75	0.67
	G5	EXP1	68.8	0.37	0.63	0.72

Table 4.15: The gading performance of spatial features with SVM, RF and BDT

The best grading came from the BDT, as shown in Table 4.15, which has clearly outperformed RF and SVM. Referring to the result in Chapter 4.2.4, clusters used for this feature for grade G1 are from EXP2 while grade G3 and grade G5 are from EXP1.

The BDT built from spatial features to grade G1 are shown as appendix D.4. Note that the root node starts with number of  $C5_m$  level (which means cluster C5 in middle distance from the tissue surface), while the tree-G1 model from texture features also starts with cluster C5. BDT for grading G3 with spatial features are shown as appendix D.5 while BDT for grade G5 is as D.6. The root node for the BDT-G3 starts with a number of cluster  $C1_d$  (which located far from the epithelial layer). This tree is not balanced, similar with the tree-G3 model built from texture feature.

Texture features of the whole annotated regions were also investigated and reported in Chapter 4.2. Best parameter settings at the pixel level was clustering patches into 5 clusters, with patches sizes  $100\text{pixels} \times 100\text{pixels}$  across annotated regions at 20X, with GLCM in four directions among ten neighbouring pixels. The CCM features on the patch level were calculated for 4 neighbouring clusters in all four directions. Binary classification were carried out using BDT and the best feature for grading G1 was the frequency of clusters from EXP2 (without correlation at both pixels and patch level).

G3 classification was best using the contrast, energy and homogeneity among clusters in CCI of EXP1 (without correlation features at patch-level) while G5 was best classified with the contrast, correlation, energy and homogeneity among clusters in the CCI images of EXP1 as well. Initially, we reported that annotated regions with grade G5 has return many 'NaN' values for its correlation features in pixels-level. The result shows that the correlation features at pixel-level are worth investigating as it can be an indicator for grade G5 of annotated regions.

Then, the spatial features of clusters, in relation to the epithelial layer were carried

out as well. Texture features selected are similar with the previous steps, and the grading result is much higher with the BDT, compared to SVM and RF. The BDT classification models from both the whole tissue and spatial features were selected to be implemented in the whole virtual slides grading; in the next chapter.

## 4.4 Conclusions

Grading dysplasia in BO is tricky as there are many common features shared in the series of changes from Barrett's with no dysplasia, to Barrett's with severe dysplasia. The experiments to investigate the significance of epithelial layer texture features have been reported in Chapter 4.1. The result suggested that we accept the  $H_0$  using the tissue boundary as the reference line. Co-occurrence clusters of patches in one direction ( $0^\circ$ ) at 10X magnification along the tissue boundary give the best result (82.5% AP) to differentiate between dysplasia and non-dysplasia regions. Patches were of size 100pixels width  $\times$  150pixels height perpendicular to the normal surface. This knowledge could be applied prior to grading the regions created, shown in the next chapter It could filter out filter out non-dysplastic regions from further processes thus, potentially saving time and computational burdens.

This chapter has demonstrated that our approach of creating matrices based on clustered patch co-occurrence has the ability to differentiate tissue textures that co-exist in the same region. Compared to the original GLCM, CCM works on higher image magnification and enables the measurement of the spatial arrangement of tissue types. This can be used to grade dysplasia in BO.

Another advancement of the existing method is in solving the 'border effect'. Instead of leaving tissues around their epithelial layers off the tissue analysis process, patches were taken from the border automatically. The location of some tissue structures with reference to the epithelial layer also contains important patterns to enable region analysis and grading with machine learning techniques. The result achieved by investigating textural features in epithelial and lamina separately in annotated regions has reached better consensus than that achieved by existing pathologists.

Most importantly, our novel contributions towards pathologist society is that we have demonstrated that morphological changes in BO dysplastic tissue can be measured and it maybe possible to standardise the grading itself. The grading rules from the BDT could be used to explain the reasoning behind each decision as each cluster represents a certain texture.

The next steps of this research is to embed the region creation method as discussed

in Chapter 3, as well as the BDT grading models validated in this chapter, to grade the whole virtual slides. A suitable decision support mechanism which considers the grading decision as well as the supportive evidence will be applied to achieve consensus grading for the whole virtual pathology slide analysis.

# Chapter 5

## Diagnosing whole virtual slides

---

In the previous chapters, our modelling was based on regions of tissue, which was only portions of tissues, while the tissue itself is only a portion of the whole virtual slide. Thus, the images content, memory size and variability are much smaller than the whole virtual slides. In this chapter, however, these models are implemented on whole virtual slides, which have multiple numbers of tissues and indirectly multiply the number of regions.

Applying the models that we have developed and trained on a graded annotated region is more challenging as we are implementing them on a larger scale. This is because the processing time and cost, as well as the number of tissues, regions, patches and grades along with noises and artefacts will be multiplied. Furthermore, the challenge to achieve a consensus agreement on grading dysplasia will increase as well. To understand this, the standard procedures to grade dysplasia for pathology glass slides is revised. As discussed in Chapter 1.1, BO pathology slides contain several pieces of tissue which are examined by the pathologist(s). Some of them are annotated and graded but usually grades are varied between G1 to G4 or between G3 to G6. As a standard procedure, pathologists have to take every annotated region into account before deciding the final grade of dysplasia for the whole pathology glass slide. Naturally, having more regions graded in a whole virtual slide will multiply the grades contradiction and confusion.

Referring to the processes in Figure 5.1, this chapter encapsulates our works in Chapter 3 as virtual slide dis-integration modules. Specifically, we integrate the implementation of tissue detection, noise removal (as presented in 3.2 and 3.3) and region creations and selection from tissue detected (as in 3.4) on the whole virtual pathology slides. Then,

feature extractions and clustering in the epithelial layer, and lamina as well as the spatial information will be carried out according to the parameters selected in chapter 4; on each of the regions selected. Then, the regions will be graded with the BDT models selected previously (in chapter 4).

The feature selection, clustering as well as the grading carried out will then be implemented as the input to a virtual slide re-integration processes. Then, finding the consensus grading for each of the virtual slide according to grades suggested on each regions is the main challenge to be discussed, investigated and acted upon, in this chapter.

The contribution of this chapter lays in combining the images and grades of each region into their respective slides, as well as the method to achieve consensus grading between all the BDT models for each region. A consensus grade among the BDTs will be formulated to grade the whole virtual slides as either dysplasia with grade G1[0,1], G3[0,1] or G5[0,1].

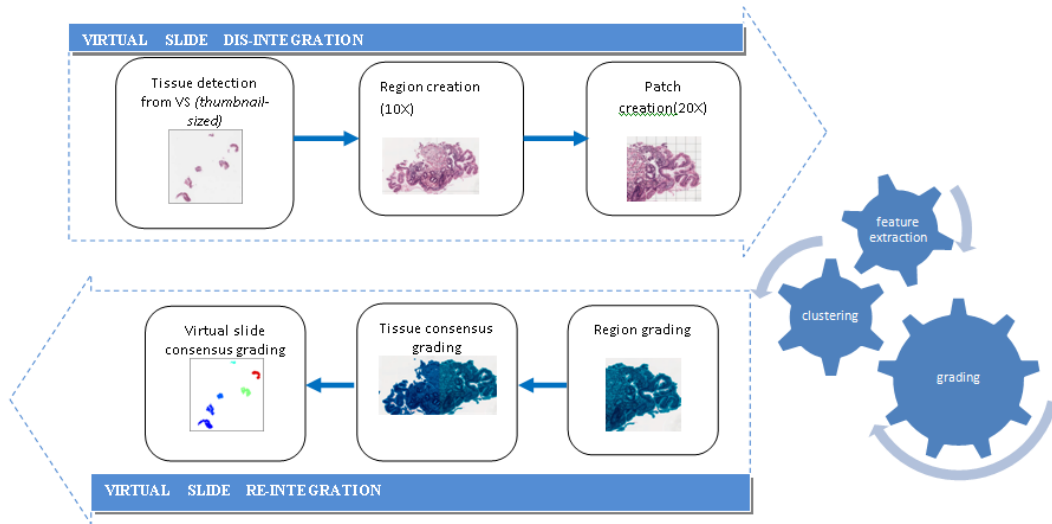


Figure 5.1: Processs involved in grading a virtual slide.

## 5.1 Virtual slides preparation

Referring to Chapter 3.1, Table 3.2, 60 whole virtual slides were used as a training set, where annotated regions in each of them were used for training on validation of tissue texture clustering and region grading. Another 15 whole virtual slides consisting of five for G1, five for G3 and another five for G5 were set aside for testing purposes.

These virtual slides have been viewed by two expert pathologists, and the same reviewer has independently reviewed the corresponding glass slides as well. As detailed



in Table 5.1, the diagnosis between two experts on glass slides has managed to achieve a very good agreement. But the virtual slides diagnosis achieve only 80% agreement with three cases concluded with 'no consensus' (NC): slide ID 11063, 10857 and 11054. One other case was diagnosed as 1 grade lower than the glass diagnosis: slide ID 10790. All four cases were originally grade G3 (glass slide diagnosis), which is the most fuzzy grade between the three grades. However, we will compare our suggested final grading for the virtual slides with the pathologist grading on the glass slides.

slideID	virtual diagnosis			glass diagnosis		
	expertA	expertB	consensus	expertA	expertB	consensus
13348	5	6	G5	5	6	G5
11040	1	1	G1	1	1	G1
13154	5	6	G5	5	6	G5
11063	3	2	NC	4	3	G3
10586	1	1	G1	1	1	G1
13083	5	6	G5	5	6	G5
10857	1	3	NC	4	3	G3
11014	1	1	G1	2	1	G1
10790	2	1	G1	4	3	G3
10829	1	1	G1	1	1	G1
11035	1	1	G1	1	1	G1
13303	5	6	G5	5	6	G5
11054	3	1	NC	4	3	G3
13239	6	6	G5	6	6	G5
11013	4	4	G3	4	4	G3

Table 5.1: List of virtual slides test data with the consensus achieved for glass and virtual slides diagnosis.

## 5.2 Tissue and region selection

Implementing tissue detection and noise removal as explained in Chapter 3 has given us 3051 regions for further analysis. These regions came from 242 pieces of tissues detected and selected automatically from these 15 virtual slides.

Regions with the average grey level of  $<0.73$  or the entropy  $<6.3$  as suggested by experiments in Chapter 3 is accepted as candidates tissue. In addition, wax smears were eliminated by excluding regions where the sum of grey level histogram between bin 190 and 210 is higher than the overall mean histogram.

To ensure only regions containing epithelial layer go to the next process, regions

around the bounding box of the detected pieces of tissue will be selected (as shown in Figure 3.10). Then, the last additional criteria is to choose regions with average grey level value  $>0.75$  and total white pixels (background) is  $<35\%$  of the whole region. All these threshold values for filtering purposes has been explained in detail in Chapter 3.4.2.

Table 5.2 shows number of detected objects from one virtual slide. Each object is then further analysed at 5X magnification level before it is accepted as a candidate tissue. Each candidate tissue is then divided into regions at 20X magnifications. Then, each region will be analysed with its basic property such as the entropy and mean grey level value, as well as number of background pixels as stated above. Based on this information, meaningful regions were accepted for further tissue extraction process.

Tissue number	regions created	regions accepted
tis1	0	0
tis2	23	5
tis3	24	3
tis4	15	6
tis5	16	6
tis6	0	0
tis7	32	8
tis8	20	7
tis9	30	6
tis10	30	3
tis11	24	3
tis12	25	7
tis13	4	2
tis14	28	5
tis15	25	6

Table 5.2: Table show sample of tissues detected and filtered, as well as the regions accepted for region grading process from a virtual slide using our selected parameters and threshold values.

### 5.3 Implementation of epithelial layer analysis

Out of 3051 regions created from all the accepted tissue, only 1735 regions were accepted for epithelial layer, texture and spatial features layer analysis. The selected region contains background where tissue boundary detection was carried out. The detection process aims to ensure that the correct boundaries are detected, and not muscularis mucosa.

Then the epithelial layer tissue analysis is implemented. This is to select only dys-

plastic tissue for further processing, thus reducing the computational cost in processing normal tissues. Referring to Chapter 4.1, the final result of this process is to grade epithelial layer in a region into dysplastic or non-dysplastic tissues.

In order to do this, un-rotated patches along the detected boundary from these regions were analysed and clustered based on the clustering model selected. Patches were clustered into 5 clusters with 78% of the regions (1354 regions) being classified as dysplastic tissue using the cluster co-occurrence. These regions will be used for the next processes.

## 5.4 Implementation of BDT models

Implementing the clustering model as well as the BDT models on the CCM features was straightforward, but the feature extraction took a long processing time with an average of one virtual slide taking five working days. Therefore, we used several processing nodes on the large-scale advanced research computing (ARC1) provided by the university to run a parallel feature extraction and clustering for all these regions.

CCM texture features of each region of these test sets were extracted as explained in 4.2.2 previously. These texture features were used to grade the regions into G1[0,1], G3[0,1] and G5[0,1] using the BDT models selected in Chapter 4.2.3.

Therefore, to achieve a consensus grading for each region, a positivity table is used to count in the support value gained from every model. A positivity table lists all grades given by the BDT models for each region. Grade which do not conflict with the other are counted as support. In a positivity table, a grade is considered strong if only one BDT model produces a positive grade (G1[1] or G3[1] or G5[1]), and the other model gives a negative grade (G1[0] or G3[0] or G5[0]). On the other hand, if all three BDTs produce a positive grade, or all negative grades, no consensus grade is achieved.

The score from the positivity table and the model weight is calculated to give a grading score. Table 5.3 shows the details meaning with the scoring value to show the strength of each grade for all possible combination of grades by the three BDT models.

The most probable grading out of these three models is determined by incorporating the probability holding the initial knowledge about each grade. Thus, we can use Bayes Theorem to directly calculate them. Referring to Equation ?? of Bayes Theorem, we would like to know the probability of a virtual slide being in certain grade (G1, G3 or G5)  $P(h|D_{G1,G3orG5})$ . Furthermore, the maximum probable hypothesis (called a textit maximum a posteriori (MAP) hypothesis) for a test case to fall under certain grade with a

tree-G1	tree-G3	tree-G5	score	meaning
1	0	0	2	strongly G1 and not G3 or G5
1	1	0	1	either G1 or G3, but not G5
1	1	1	0	no consensus
0	1	1	1	either G3 or G5 but not G1
0	0	1	2	strongly G5 and not G3 or G1
0	1	0	2	strongly G3 and not G1 or G5
0	0	0	0	no consensus
1	0	1	1	either G1 or G5 but not G3

Table 5.3: Positivity table and the score used to find a consensus diagnosis for a region.

known prior probability can be determined as Equation 5.1.

$$\begin{aligned}
 h_{MAP} &\equiv \operatorname{argmax}_{h \in H} P(h|D) \\
 &= \operatorname{argmax}_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\
 &= \operatorname{argmax}_{h \in H} P(D|h)P(h)
 \end{aligned}$$

The term  $P(D)$  is dropped from the final step above as it is a constant independent of  $h$ .

The  $h_{MAP}$  calculation for each grade (G1[0,1], G3[01,] and G5[0,1] with confusion matrix of  $\begin{pmatrix} 8 & 0 \\ 2 & 6 \end{pmatrix}$ ,  $\begin{pmatrix} 6 & 2 \\ 2 & 6 \end{pmatrix}$  and  $\begin{pmatrix} 6 & 2 \\ 1 & 7 \end{pmatrix}$  respectively are normalised to 0.80, 0.75 and 0.86. These maximum probabilities will be used as the  $w$  to grade regions. These confusion matrices produced by each BDT tree were shown originally in Table 4.12.

To explain this, the positivity table and score gained for regions accepted from tissue 2 of slide 13348 is shown in Table 5.4 and Equation 5.2.

The formula to calculate the overall grading score ( $GS$ ):

$$S_i = \sum w_i \times score_i \quad (5.2)$$

where  $i$  is the dysplasia grade [G1, G3 or G5]

Therefore, the  $GS_{G1}$  for the sample tissue is 0.80,  $GS_{G3}$  is 3.75 and  $GS_{G5}$  is 0.86. Based on the highest  $GS$  value, this tissue is graded as G3.

## 5.5 Virtual slides grading

Grading a virtual slide is another tricky process, as the frequency of grades occurs in a slide which is not necessarily used by pathologists. In current practice, priority was given to the most severe grade of dysplasia. For example, if there are occurrences of regions

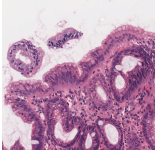
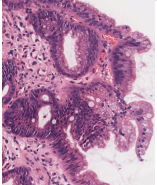
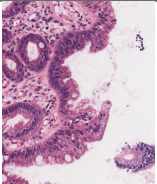
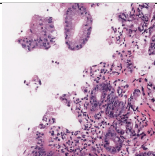
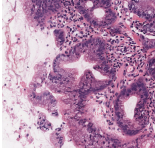
image	Region \ score	2	1	0
	tis2-1	G3		
	tis2-2	G3		
	tis2-3		G3&G1	
	tis2-4			no consensus
	tis1-5		G5	

Table 5.4: Grading score gained for regions in tissue number 2 of slide 13348.

graded as G5, the whole virtual slides will be graded as G5. This applied regardless of high frequencies of G1 and G3 found in the same slide. The same rule applies for a lower grade of dysplasia when the highest grade does not exist. However, the number of acceptable occurrence (namely  $X$ ) of G5 and G3 is not known.

Therefore, we want to calculate the ( $X$ ). In order to do this, we introduce a threshold frequency ( $F_t$ ) for G5 and G3. These thresholds are proportions of the total frequency of regions with G1, G3 and G5 detected in each class. Incorporating the current practice of grading dysplasia, the most severe grade of dysplasia detected in regions of virtual slides will be considered as the candidate grade for the virtual slide.

However, information regarding the proportion of every grade is taken into account. Referring to the second column of Table 5.7, numbers of regions graded as G1, G3 and G5 were populated according to the grading of the virtual slides to find the frequency proportion. Cases where regions are detected with G5 were  $\geq 40\%$  of the total number

of regions, so the whole virtual slide is graded as G5. Otherwise, if the number of detected regions with G3 was  $\geq 32\%$ , it is graded as G3. The frequency thresholds or the proportion are shown in Table 5.5.

Region \ Slide	G3	G5
G3	0.333	0.287
G5	0.048	0.245

Table 5.5: Proportion of regions detected with G3 and G5 for each class of virtual slides.

The occurrence of G1, G3 and G5 suggested by BDTs in each virtual slide are calculated, and the number of G3 and G5 classifications are compared against its  $f_i$  value. The result for all virtual slides test set using the threshold frequency with  $X=5$  is shown in Table 5.7. We have also included the grading suggestion if the highest frequency grading is used, for comparison. Furthermore, the confusion matrix for the virtual slides grading test set is shown as Table 5.6.

suggested \ pathologist	G1	G3	G5
G1	4	1	0
G3	1	3	0
G5	0	1	5

Table 5.6: Confusion matrix between our grading method and the ground truth (glass slide diagnosis).

Based on the table, we have managed to achieve a high agreement score of 0.80 KV with the glass slide grading by the pathologists with the implementation of a threshold frequency. This agreement score is a massive improvement compared to the agreement score between the glass slide grading with the highest frequency grading, which is only 0.47 KV. Furthermore, high true positive results for all three grades shows that our three classes binary grading models are reliable.

SlideID	Grade of tissues accepted	$f_{G1}$	$f_{G3}$	$F_t$	$f_{G3}$	$f_{G5}$	$F_t$	$G5$	D	E	F	G
13348	G5, G5, G3, G1, G3, G5, G3, G1, G1, G5	3	3	4	7	3			G5	G5	G5	G5
11040	G1, G1, G3, G1, G1, G3, NC, G1, G1, G1, G1, G3, G5, G3, G1	11	4	6	1	4			G1	G1	G1	G1
13154	G1, G1, G5, G5, G3, G5, G3, G5, G5, G5, G3, G1, G3, G5, G3, G3, G5	1	6	7	11	5			G5	G5	G5	G5
11063	G1, G1, G1, G5	3	0	1	1	1			G5	G3	NC	G1
10586	G1, G1, G3, G1, G3, G1, G1, G1, G3, G3, G1, G3	7	5	4	0	3			G3	G1	G1	G1
13083	NC, G1, G3, G5, G5, G3, G3, G5, G1, G5, G5, G3, G3, G1, G5, G5, G3, G3, G1, G3, G3, G3, G3, G3	4	12	8	6	6			G5	G5	G5	G3
10857	G1, G3, G3, G1, NC, G1	3	2	2	0	2			G3	G3	NC	G1
11014	G1, G1, G1, NC, G3, G1, G1, G3, G5, G1, G5, G3, G1, NC, G1, G3	8	4	5	1	4			G1	G1	G1	G1
10790	G3, G1, G1, G1, G1, G3, G3, G1, G3, G1, G1, G1, G5, G1	9	4	5	1	4			G1	G3	G1	G1
10829	G1, G1, G3, G3, G1, G1, G1, G3, G1, NC, NC, G3, G1, G1, G1	9	4	5	0	4			G1	G1	G1	G1
11035	G5, G5, G3, G3, G3, G1, G1, G1, G1, G3, G1, G1, G1	8	4	5	2	4			G1	G1	G1	G1
13303	G1, G1, G3, G5, G5, G1, NC, G3, G3, G3, G5, G5, G5, G1, G1, G1	7	3	5	5	4			G5	G5	G5	G5
11054	G3, G1, G1, G1, G3, NC	3	2	2	0	2			G3	G3	NC	G1
13239	G3, G3, G5, G5, G1, G1, G1, G5, G1, G1, G1, G3	6	3	4	2	3			G5	G5	G5	G1
11013	G1, G1, G1, G3, G1, G1, G1, G3, G3, G1, G1, G3, G1, G1, G3, G3	10	6	5	0	4			G3	G3	G3	G1

Table 5.7: Grading result for virtual slides using our CCM features and thresholded frequency value.  $f_{G1}$  is the frequency of G1;  $f_{G3}$  is the frequency of G3;  $f_{G5}$  is the frequency of G5;  $f_t$  is the threshold frequency of each grade accordingly; D is our suggested grading with threshold frequency; E is the consensus grading by two pathologists on the glass slides; F is the consensus grading by two pathologists on virtual slides and G is suggested grading based on frequency only.

## 5.6 Conclusions

This chapter is an important part of the thesis contribution as proof that the whole process of virtual slide dis-integration for tissue detection, region creation and grading as well as re-integrating tissue into the respective virtual slides were tested and worked. The workflows can be revisited at Figure 1.2.

The tissue detection and selection process as discussed in Chapter 3 were implemented on 15 test virtual slides, giving us 222 detected tissues. These tissues were then disintegrated for texture analysis, using the region creation approach as discussed in Chapter 3.4. This process has given us 3051 regions for analysis and grading in the region layer. The implementation of tissue epithelial, texture and spatial analysis in this level has given us a variety of grading for each virtual slide. These have proven that our CCM feature, as well as the BDT models selected during analysis and experimental set up in Chapter 4, managed to model the hidden knowledge of texture relationship for grades G1, G3 and G5 of dysplasia.

Another important contribution of this chapter is the discovery of calculating the consensus grading for virtual slides diagnosis. In region level, we have tested the positivity table and scoring method on our BDT models to include supportive grading by the other decision model, indirectly avoiding conflicting grades. There are still few tissues where no consensus is gained, but the grading for the whole virtual slides are not interrupted. Then, we have measured the proportion of region's grade for virtual slides with G1, G3 and G5; and use this as threshold frequency for consensus grading of the virtual slides. The result were very good compared to the existing agreement achieved between pathologists. Therefore, the  $f_i$  can be used as acceptable occurrence  $X$  of graded region to grade dysplasia in BO.



# Chapter 6

## Conclusion and future work

---

### 6.1 Summary of work

The objective of this research is to help identify and measure dysplastic changes in BO virtual pathology slides using its textural features and spatial relationships. In order to do this, experiments were set up to extract the best representations of tissue texture and spatial pattern.

The first novel contribution of this research is the solution for ‘border effect’ suffered by BO, colon, breast, prostate and oral tissue (among others) during digital image processing. A lot of references regarding the tissue condition were linked to the surface maturity, the complex structure of the epithelial tissue and the cell’s polarity, as we can see from Table 2.2.

Therefore, when the boundaries are curvy or in a viliform structure, the morphological changes of tissue structures at the lamina propria region follow. Traditional image processing techniques usually work within a square or rectangular window frame aligned to the image frame, but the nature of dysplastic tissue of BO is its complex boundary that does not aligned to the image frame. This is the ‘border effect’ which has been avoided by many researchers who works with tissue images by manually selecting sample tissue without a surface membrane [23, 55, 60].

Therefore, we have develop a method to extract and analyse tissue texture along the epithelial layer. This method has proven to be successful as it achieved 82.5 AP with 0.82

precision and 0.86 recall values to grade regions into dysplastic or non dysplastic. The approach has provide a solution to overcome the 'border effect' problem and could easily be adapted to other images. In addition, it enables texture-based analysis to be used where tissue architecture is mainly used. The study on tissue on the surface membrane has been attempted before on rat brain hippocampal tissue images in [32] using architecture-based approach.

Our second contribution is the texture-mapping method in producing the Cluster-coded Co-occurrence Image(CCI). The co-existence of many types of tissue texture and its ratio within a tissue sample (or a region) has been a good indicator of tissue condition. However, there is no standard map or atlas available for our domain, unlike the whole brain architecture or prostate cancer grading. Realising that the morphological changes in BO involves many types of tissue textures that co-exist within a region, we have developed a new texture mapping technique.

In order to do this, patches were created across regions of tissue and clustered based on texture similarity. The unsupervised clustering model used was trained with tens of thousands patches from BO tissue. The patches were then re-assigned into their respective locations to produce the CCI. This approach has been publish in [3] and gained verbal approval from several pathologist who have looked at the images.

Co-occurrence features among these clusters within a region were calculated to produce the Cluster Co-occurrence Matrix (CCM).The CCM represents the relationship between types of tissue texture within a window frame, which is at a higher level from the existing GLCM approach, which looks at relationship between pixels. The CCM feature has been used to train a binary decision tree, to grade dysplasia into G1, G3 or G5. These trees were used as a decision model on the test sets images, and has achieved 87.5, 75.0 and 81.3 AP with KV of 0.75, 0.5 and 0.63. Using the above experiments, the use of CCI images has been shown to provide a statistically significant improvement and has also been published in [3].

Our third novel finding is the understanding of spatial arrangement of tissue texture types with reference to the epithelial layer. However, the grading performance from these feature is not as good as the CCM features with 75.0, 68.8 and 68.8 AP with KV of 0.5, 0.37 and 0.37 achieved.

Finally, the grading models were applied on the whole BO Virtual Slides. This is another contribution to the machine learning and pathology society as previous research was to locate the dysplastic tissue. We have advanced two closely related pieces of research carried out by Hamilton et al. in [40] and Snape in [109]. Hamilton uses texture analysis to locate dysplastic tissue automatically on colorectal tissue. Snape, which has

implemented texture and architecture features using DT and colour histograms has only achieved 62.5 AP.

## 6.2 Discussion and Future Work

This research has contributed to the computer vision, image processing and machine learning community where all of these play a major role in developing a diagnosis tool by identifying patterns and rules which might be too subtle for human perception. Pathologists especially, can benefit from this development to help in teaching, diagnosing and visualizing the reasoning behind each grading with quantifiable features.

The proposed models are not to be used solely for grading dysplasia, but more as an aid to assist pathologists in carrying out their work. Our tissue selection, region creation and region selection method can help pathologists to identify dysplastic regions quickly.

Our attempt to investigate the epithelial layer with a solution of the ‘border effect’ has proven to be successful as we could not simply avoid regions without boundaries in our domain. For other tissue types, knowing the depth of epithelial layer (if available) might be useful. Thus, the patch size for texture extraction is known and better accuracy can be achieved.

The Cluster Co-occurrence Image developmental method used to map the tissue texture type might also be used in other domains as well. However, a good clustering model would be required for domain-specific texture.

A major challenge of the work has been the many experiments required to find the optimum combination of parameters. Further work is required to organize these and to prove the significance of each parameter.

The outcome of this research has the potential to provide more information for pathologists to challenge or support their grading decisions. Ideally, further work should be carried out, to evaluate the grades for each self-extracted regions from Chapter 5. If we had the ground truth data for the grade given by pathologist for all 3051 regions, we could then calculate the agreement with grades suggested by our BDT models. However, this would require a significant commitment from pathologists over a longer time frame.

The final contribution of the Cluster Co-occurrence image produced will also benefit from future application.

Each virtual slide consists of a combination of many grades of dysplasia. Thus, we have implemented the positivity table and scoring method to calculate a region and virtual slides consensus grading. Therefore, we would like to see if we have managed to achieved a pathologist-like consensus diagnosis. This can be done by testing the method on other

similar diagnosis pattern such as breast or tongue cancer.

Extensive work on grading pathology slides by pathologists would be required to enable the consensus grading with supervised learning in the future. The implementation of Bayesian Network or Gaussian Classifier might be useful as it can incorporate a prior knowledge and the probability for each grade.

Due to the uniqueness of tissue texture for BO, the BDT models were not generally applicable for other type of tissue disease. Nevertheless, implementation of these models on dysplastic tissue of colon can be tested as colon tissue have similar architecture with dysplastic oesophagus tissue.

Lastly, the visualisation of each region's grading in each virtual slides can be useful. Pathologists, or other end user can then see which part of the virtual slides are considered severe and which are not. This can be represented by a colouring scheme for each grade and the colour saturation shows the confidence level of each grading.

## **Appendix A**

### **Colour normalisation result**

normalisation technique	relevance vector	regionID						
		7	55	158	290	427	612	618
NormBimodal	no barretts	OK	OK	X	X	OK	OK	X
	liver1	OK	OK	X	X	OK	OK	X
	liver2	OK	OK	X	X	OK	OK	X
	liver3p1	OK	OK	X	X	OK	OK	X
	liver3p2	OK	OK	X	X	OK	OK	X
	liver4a	OK	OK	X	X	OK	OK	X
	liver4p1	OK	OK	X	X	OK	OK	X
	liver4p2	OK	OK	X	X	OK	OK	X
	liver4p3	OK	OK	X	X	OK	OK	X
	HE	OK	OK	X	X	OK	OK	X
Reinhard Weighted	no barretts	OK	OK	OK	X	OK	OK	X
	liver1	OK	OK	OK	X	OK	OK	X
	liver2	OK	OK	OK	X	OK	OK	X
	liver3p1	OK	OK	OK	X	OK	OK	X
	liver3p2	OK	OK	OK	X	OK	OK	X
	liver4a	OK	OK	OK	X	OK	OK	X
	liver4p1	OK	OK	OK	X	OK	OK	X
	liver4p2	OK	OK	OK	X	OK	OK	X
	liver4p3	OK	OK	OK	X	OK	OK	X
	HE	OK	OK	OK	X	OK	OK	X
Reinhard Hard	no barretts	OK	OK	OK	X	OK	OK	X
	liver1	OK	OK	OK	X	OK	OK	X
	liver2	OK	OK	OK	X	OK	OK	X
	liver3p1	OK	OK	OK	X	OK	OK	X
	liver3p2	OK	OK	OK	X	OK	OK	X
	liver4a	OK	OK	OK	X	OK	OK	X
	liver4p1	OK	OK	OK	X	OK	OK	X
	liver4p2	OK	OK	OK	X	OK	OK	X
	liver4p3	OK	OK	OK	X	OK	OK	X
	HE	OK	OK	OK	X	OK	OK	X
NormReinhard		OK	OK	OK	X	OK	OK	X
RgbHist		OK	OK	OK	X	OK	OK	X

Table A.1: Normalisation output from combination of different normalisation techniques and colour classifiers.

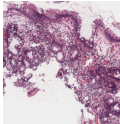
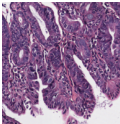
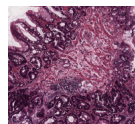
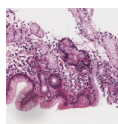
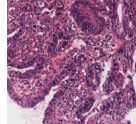
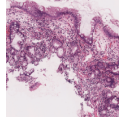
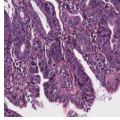
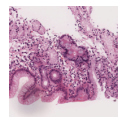
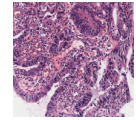
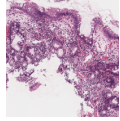
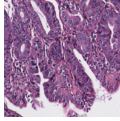
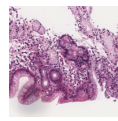
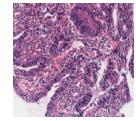
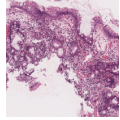
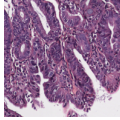
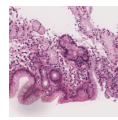
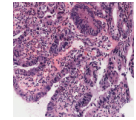
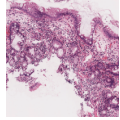
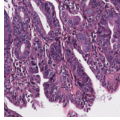
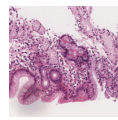
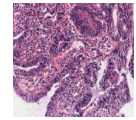
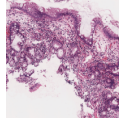
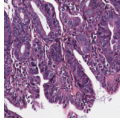
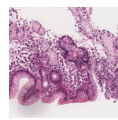
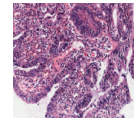
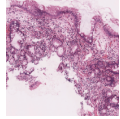
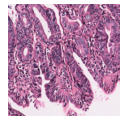
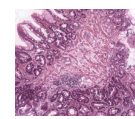
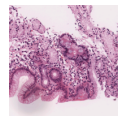
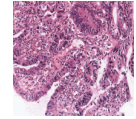
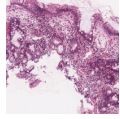
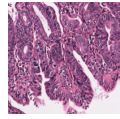
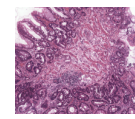
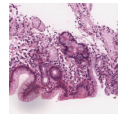
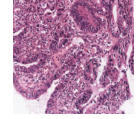
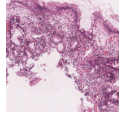
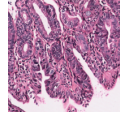
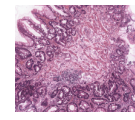
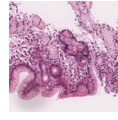
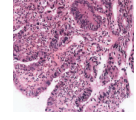
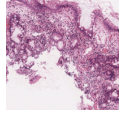
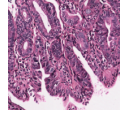
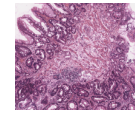
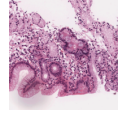
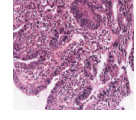
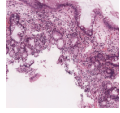
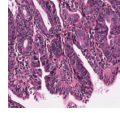
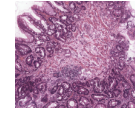
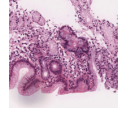
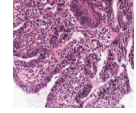
normalisation technique	relevance vector	regionID				
		7	427	158	55	612
original						
NormBimodal	HE					
	No Barretts					
	Liver1					
	Liver2					
	Liver4p1					
RVM	HE					
	No Barretts					
Reinhard	Liver1					
Weighted	Liver2					
	Liver4p1					

Table A.2: Normalised images

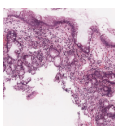
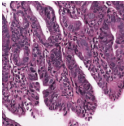
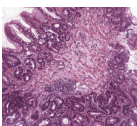
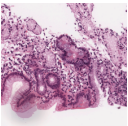
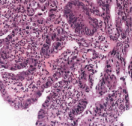
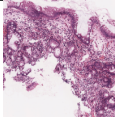
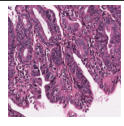
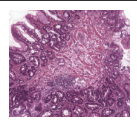
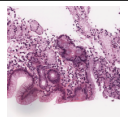
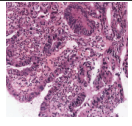
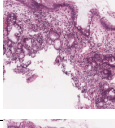
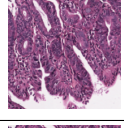
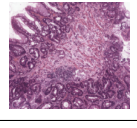
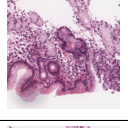
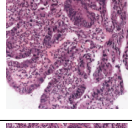
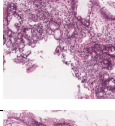
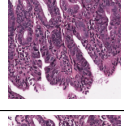
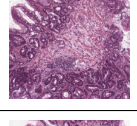
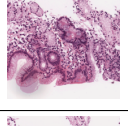
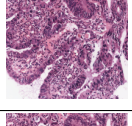
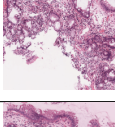
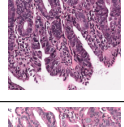
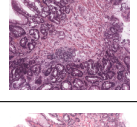
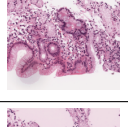
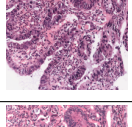
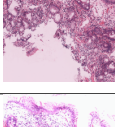
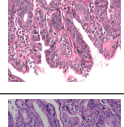
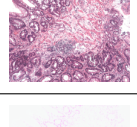
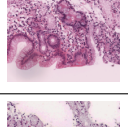
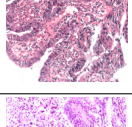
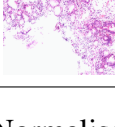
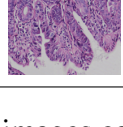

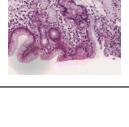
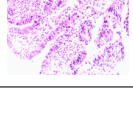
normalisation technique	colour classifier	regionID				
		7	427	158	55	612
RVM Reinhard Hard	HE					
	No Barretts					
	Liver1					
	Liver2					
	Liver4p1					
NormReinhard						
RgbHist						

Table A.3: Normalised images continued



## **Appendix B**

### **Epithelial layer analysis**

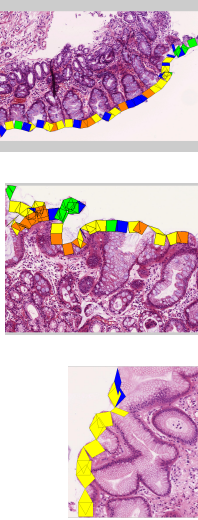
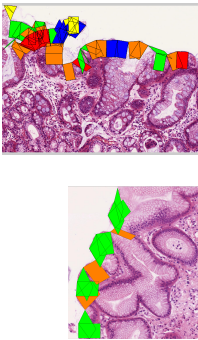
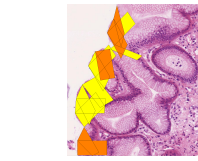
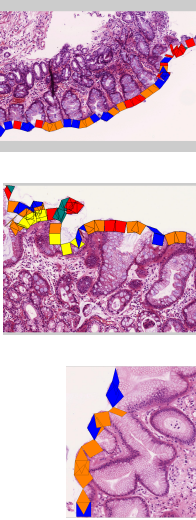
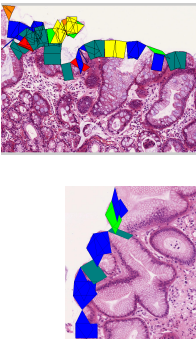
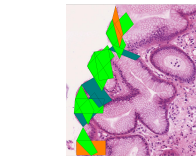
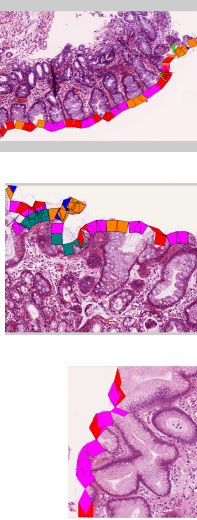
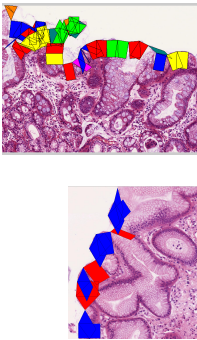
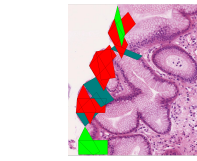
$k \backslash R$	100	150	200
5			
6			
7			

Table B.1: sample AR with clustered rotated patches on detected tissue boundary.

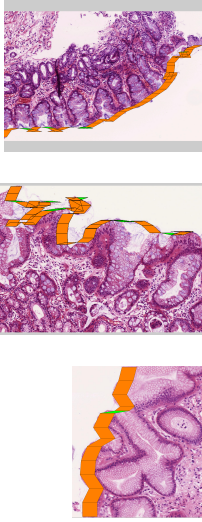
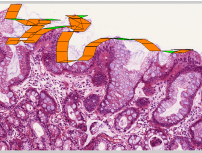
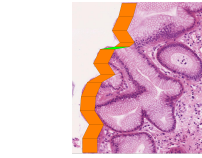
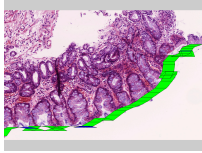
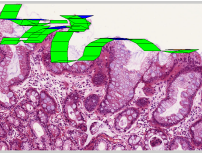
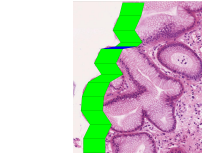
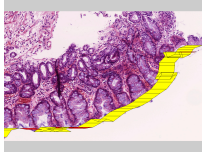
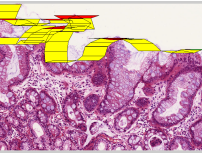

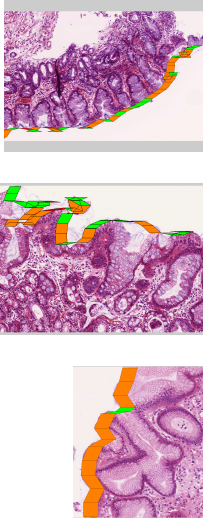
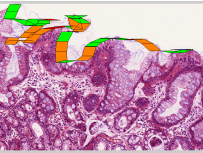
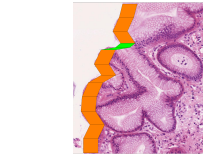
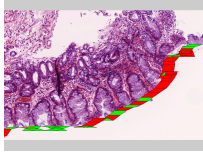
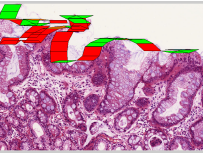
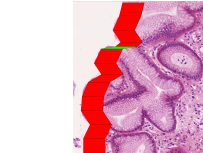
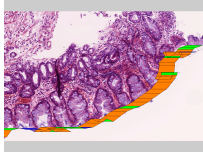
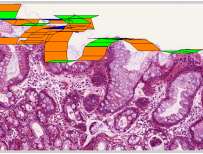
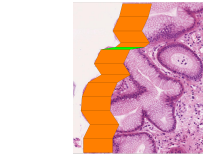
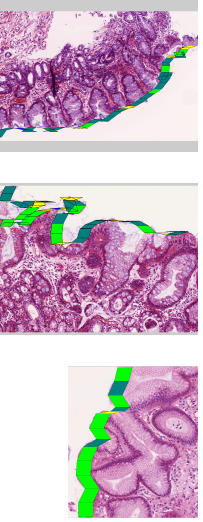
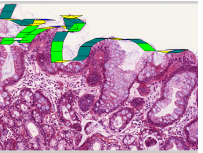
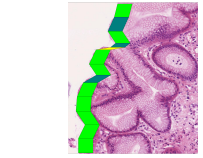
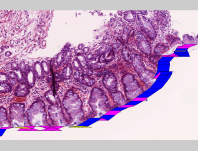
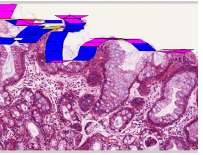
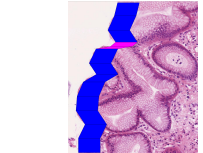
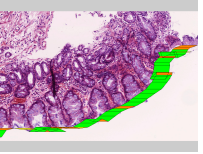
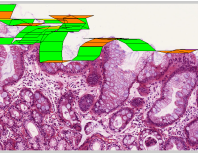
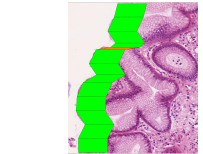
$k \backslash R$	100	150	200
5	  	  	  
6	  	  	  
7	  	  	  

Table B.2: Sample AR with clustered unrotated patches on detected tissue boundary.

## **Appendix C**

### **Texture and spatial mapping: CCI**

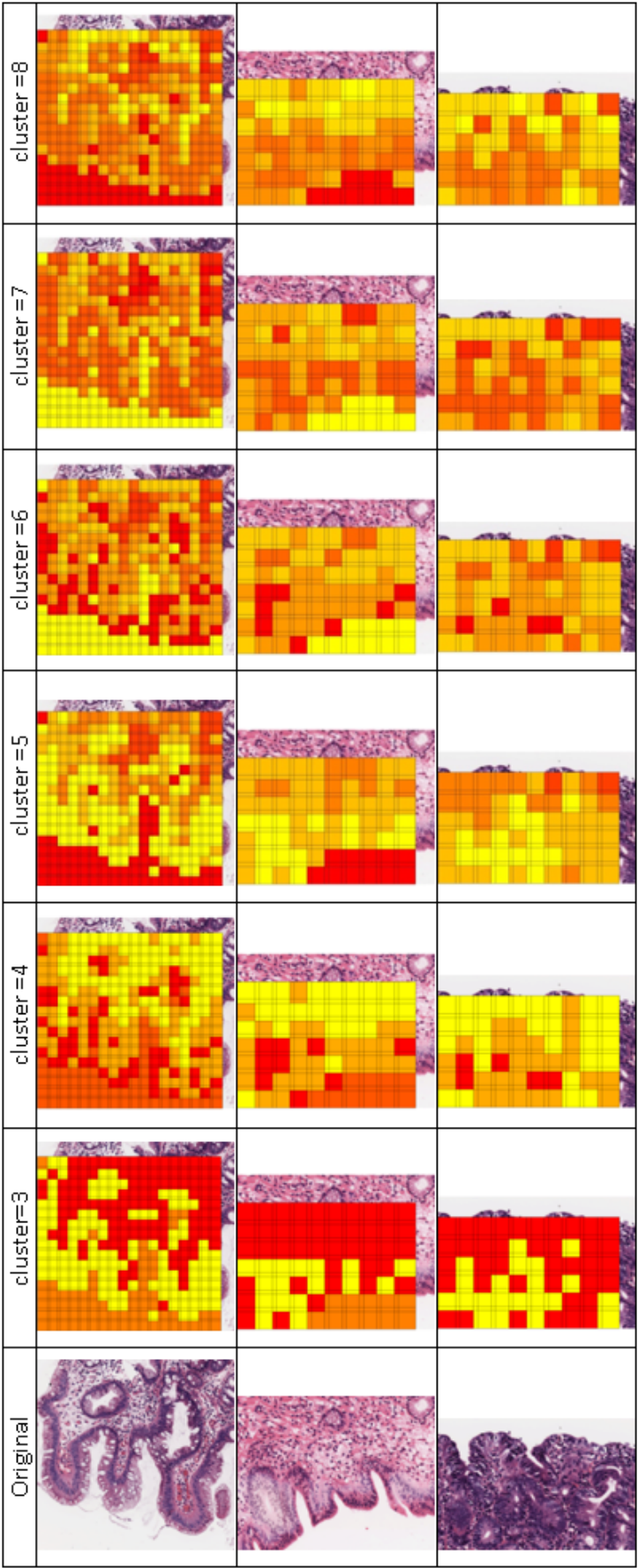


Figure C.1: Sample CCI with different  $k$  values. The zoom level is set to 20X,  $p_z=100*100$ pix and  $n=4$

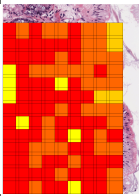
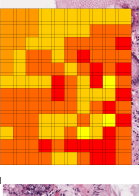
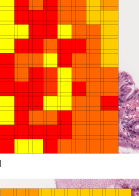
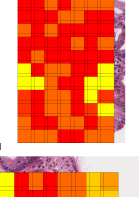
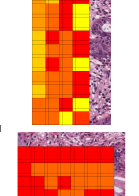
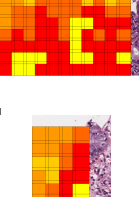
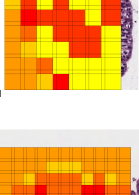
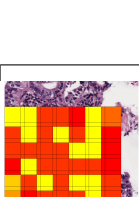
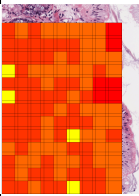
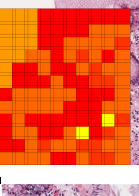
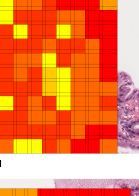
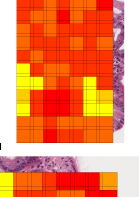
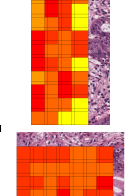
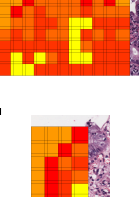
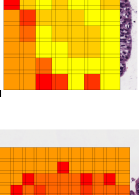
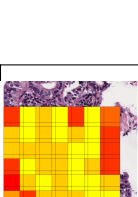
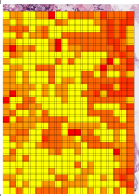
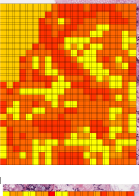
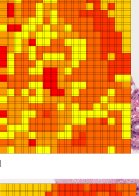
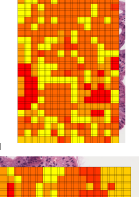
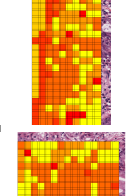
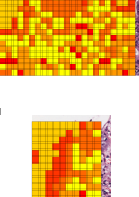
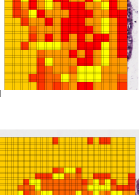
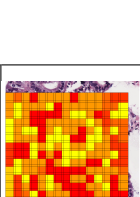
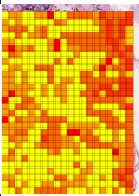
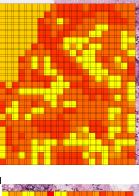
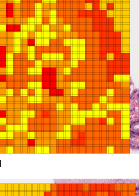
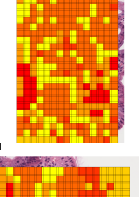
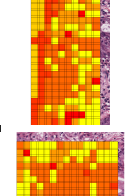
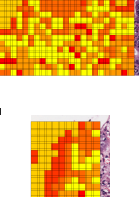
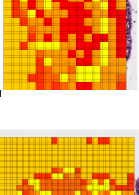
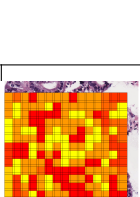
Zoom	Experiment	Sample							
20	EXP1								
	EXP2								
40	EXP1								
	EXP2								

Table C.1: CCI from EXP1 and EXP2 with different magnification.

## **Appendix D**

### **Binary Decision Trees model**



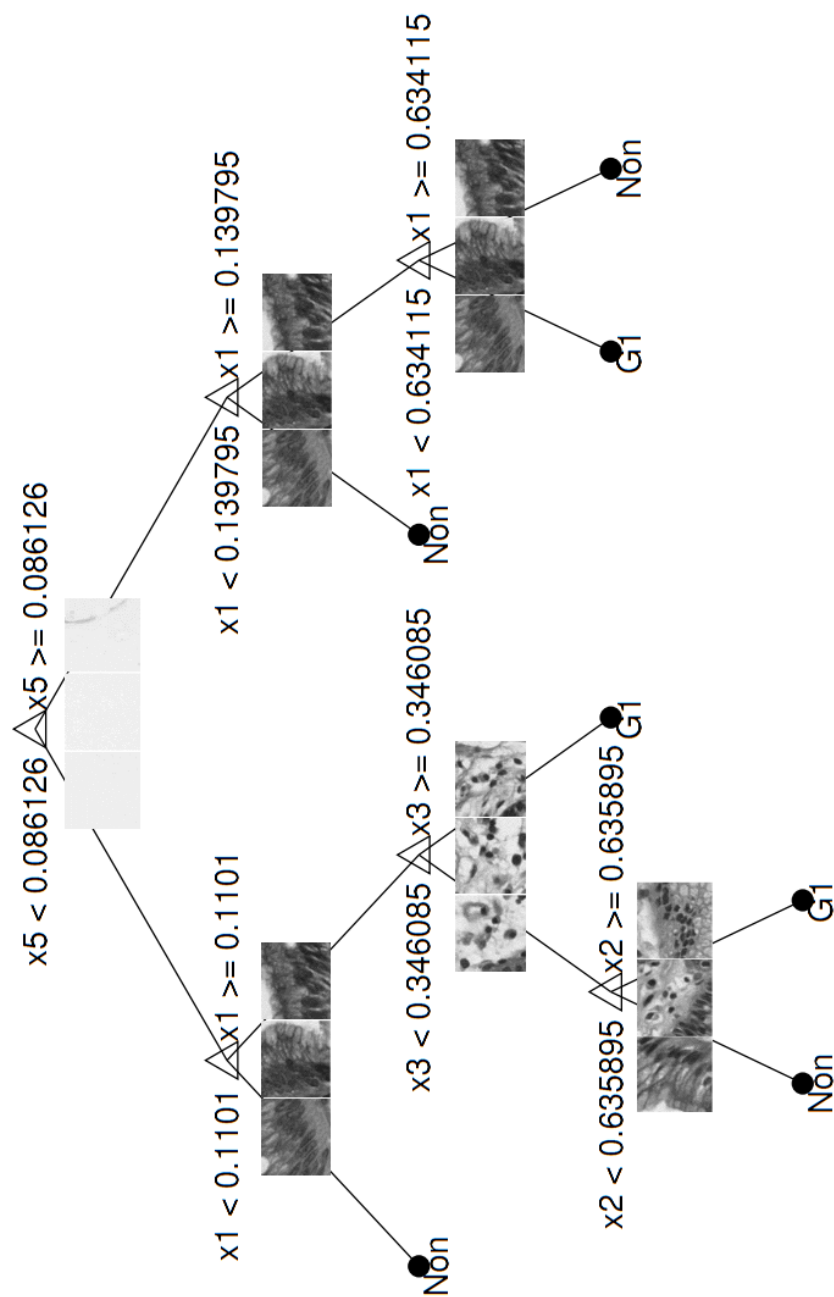


Figure D.1: BDT model selected for grading CCI into G1 or not-G1, using feature A7 from EXP2.



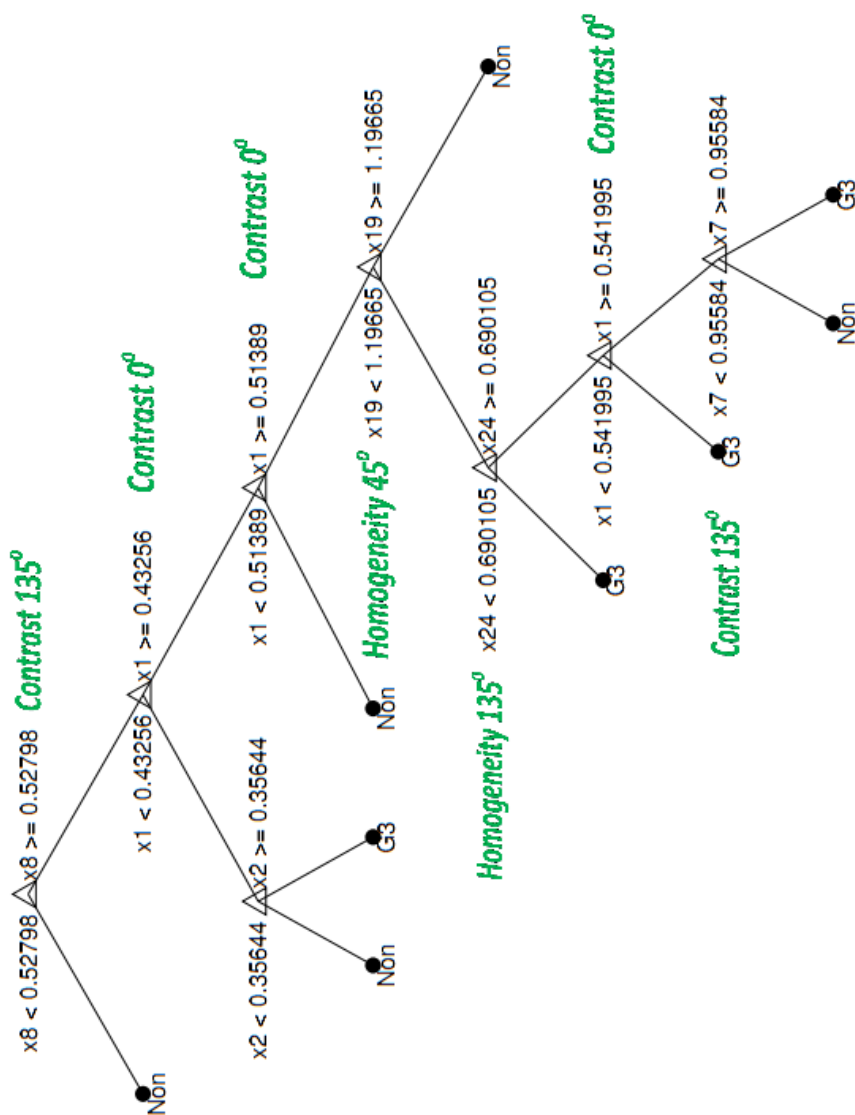


Figure D.2: BDT model selected for grading CCI into G3 or not-G3, using feature set A3 from EXP2.

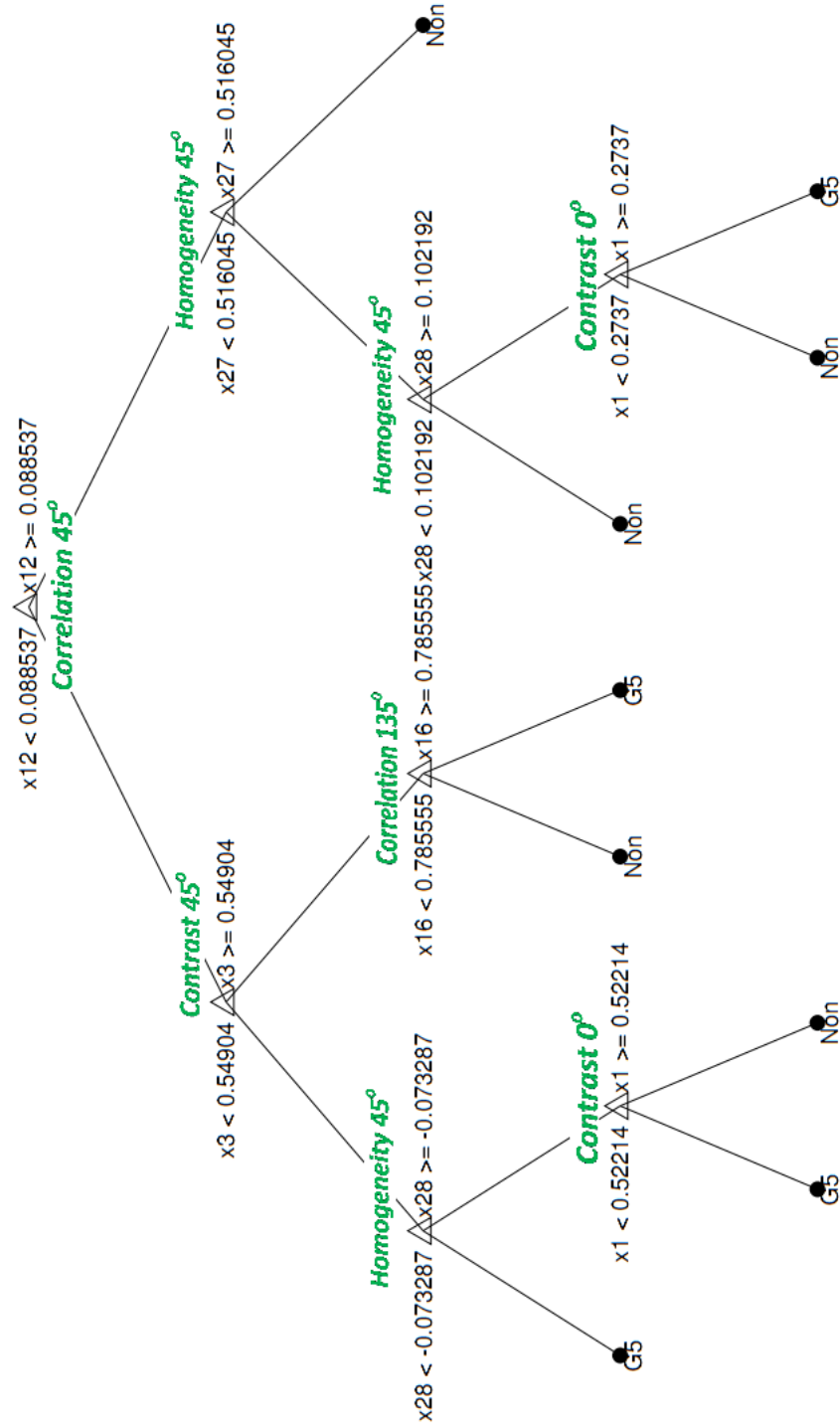


Figure D.3: BDT model selected for grading CCI into G5 or not-G5, using feature A1 from EXP1.

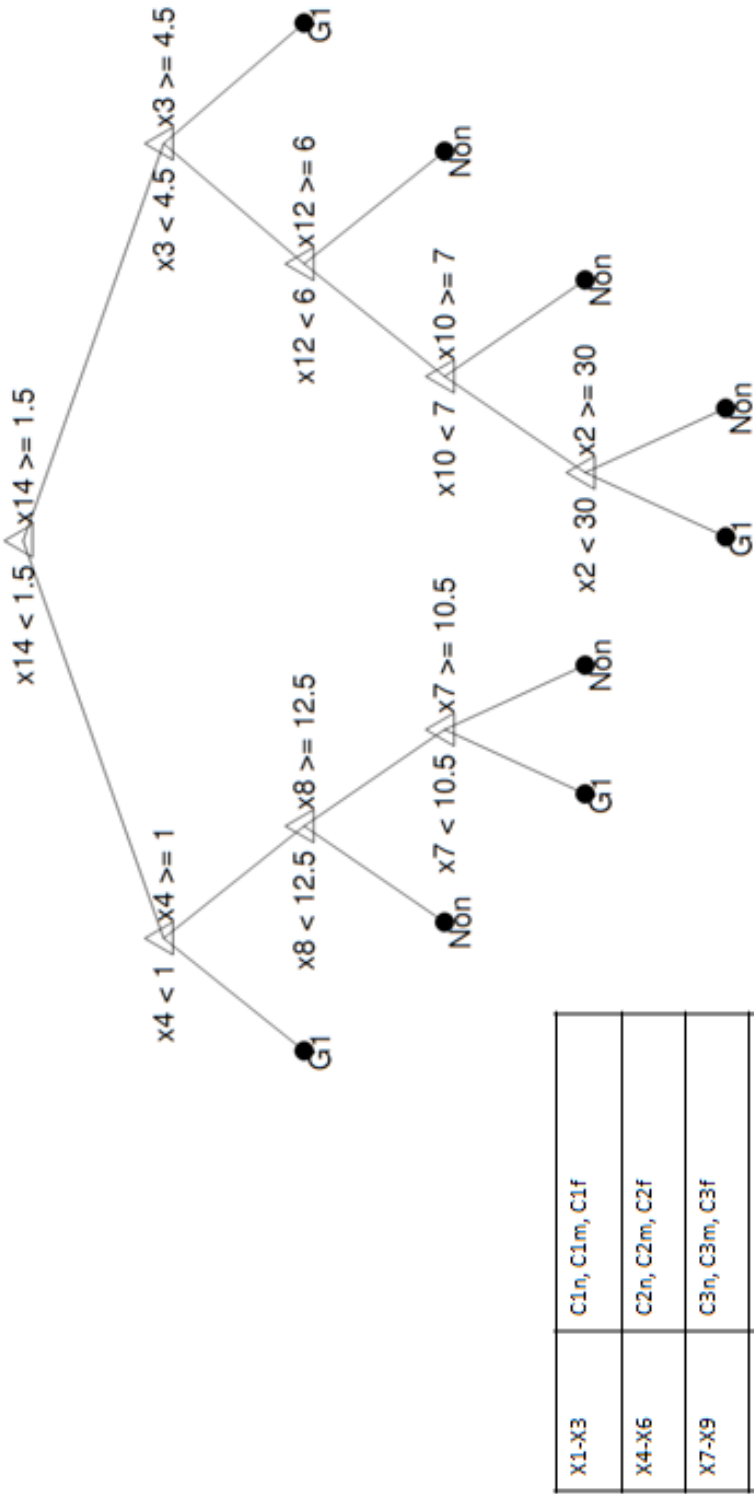


Figure D.4: Spatial BDT model for G1 vs not-G1.

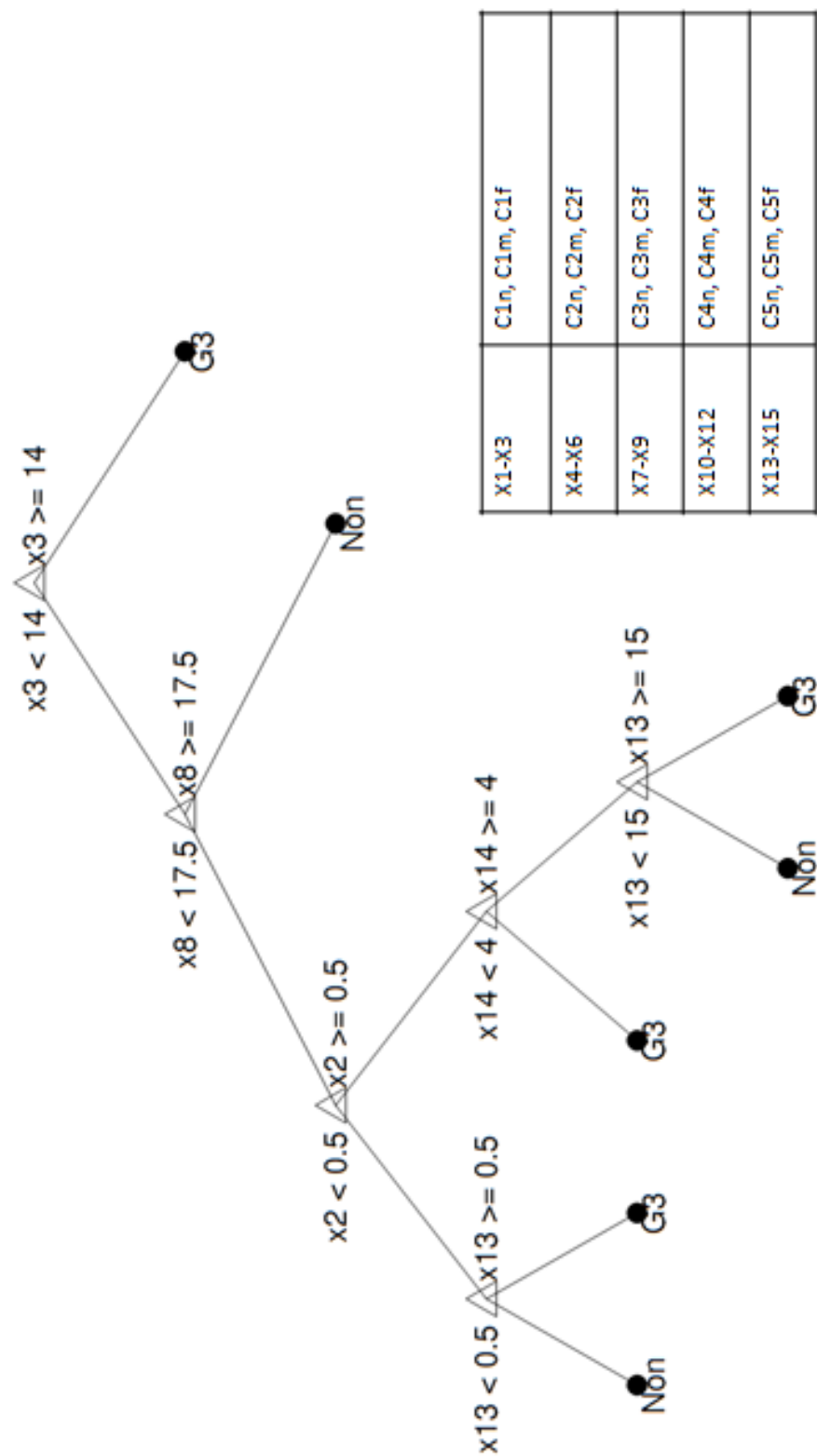


Figure D.5: Spatial BDT model for G3 vs not-G3.

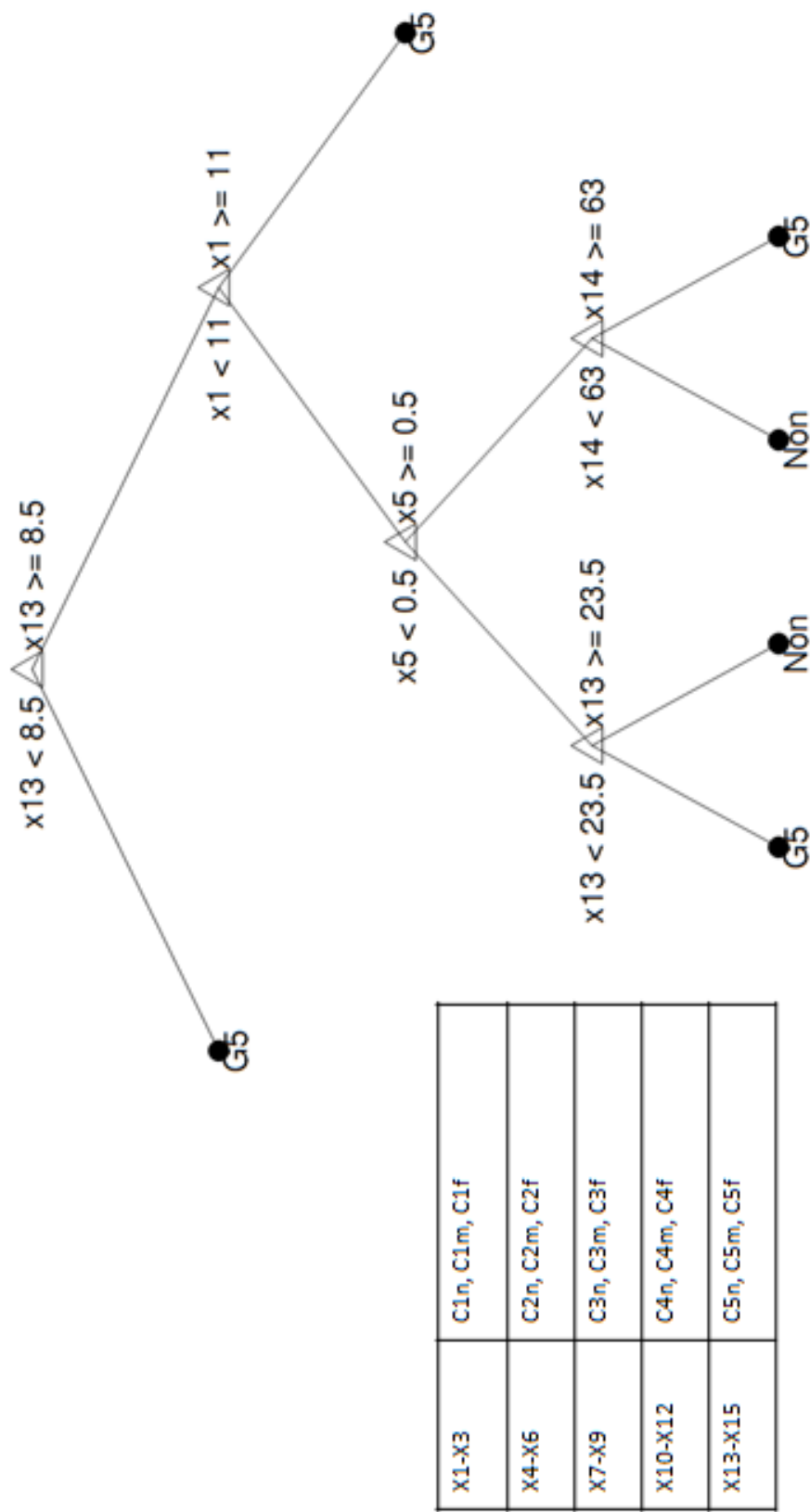


Figure D.6: Spatial BDT model for G5 vs not-G5 is not balanced and the root node starts with C5.

# Bibliography

- [1] JA. Abram. Are we making any progress in preventing barrett's-related esophageal cancer? *Therapeutic Advances in Gastroenterology*, 3(73), 2009.
- [2] A. Adam, AJ. Bulpitt, and D. Treanor. Texture analysis of virtual slides for grading dysplasia in barrett's oesophagus. In *Proc. of Medical Image Understanding and Analysis*, pages 269–273, July 2011.
- [3] A. Adam, AJ. Bulpitt, and D. Treanor. Grading dysplasia in barrett's oesophagus virtual pathology slides with cluster co-occurrence matrices. In *In Proc. of Histopathology Image Analysis: Image Computing in Digital Pathology, in conjunction with The 15th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Nice, France, October 2012.
- [4] S. Aksoy and RM. Haralick. Using texture in image similarity and retrieval. *International Workshop on Texture Analysis in Machine Vision*, 1999.
- [5] S. Alan, SL. James, and Y. Barbara. *Wheather's Basic Hitopathology: a Colour Atlas and Text*. Churchill Livingstone, 4th edition, 2002.
- [6] S. Arivazhagan, L. Ganesan, and V. Angayarkanni. Color texture classification using wavelet transform. In *Proc. of the Sixth Int. Conf. of Computational Intelligence*, pages 5–10. IEEE Computer Society, IEEE, 2005.
- [7] A. Basavanhally, S. Agnerm, G. Alexem, et al. Manifold learning with graph based features for identifying extend of lymphocytic infiltration from high grade breast cancer histology. In *Workshop on Microscopic Image Analysis with Application in Biology*, 2008.
- [8] C. Behrencruch, S. Petroudi, and S. Bond. Image filtering techniques for medical image post-processing: an overview. *Radiology*, 77:126–132, 2004.

- [9] S. Brand, TD. Wang, KT. Schomacker, et al. Detection of high-grade dysplasia in barrett's esophagus by spectroscopy measurement of 5-aminolevulinic acid-induced protoporphyrin ix fluorescence. *Gastrointestinal Endoscopy*, 56(4):479–487, 2002.
- [10] L. Breiman. *Machine Learning*, volume 45, chapter Random Forest. Kluwer Academic Publisher, 2001.
- [11] J. Bridges. Autoated detection of tumour, stroma and necrosis in colorectal cancer. Master's thesis, School of Computing, University of Leeds, 2007/2008.
- [12] J. Burthem, M. Brereton, J. Ardern, et al. The use of digital 'virtual slides' in the quality assessment of haematological morphology: Results of pilot exercise involving uk neqas(h) participants. *British Journal of Haematology*, 130(2):293–296, July 2005.
- [13] J. Caicedo, A. Cruz, and F. Gonzalez. Histopathology image classification using bag of features and kernel functions. *Artificial Intelligence in Medicine*, pages 126–135, 2009.
- [14] O. Chapelle, B. Scholkoph, and A. Zien. *Semi-supervised learning*. The MIT Press, 2006.
- [15] P. Chomphuwiset, DR. Magee, R. Boyle, et al. Nucleus classication and bile duct detection in liver histology. In *Proc. MICCAI Workshop on Machine Learning in Medical Imaging*, 2010.
- [16] M. Conio, G. Lapertosa, S. Bianchi, et al. Barrett's esophagus: an update. *Clinical Review in Oncology Hematology*, 46(2003):187–206, September 2002.
- [17] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005.
- [18] Hal Daumé. *A Course in Machine Learning*.
- [19] FR. Dee. Virtual microscopy in pathology education. *Human Pathology*, 40:1112–1121, April 2009.
- [20] FR. Dee, JM. Lehman, D. Consoer, et al. Implementation of virtual microscope slides in the annual pathobiology of cancer workshop laboratory. *Human Pathology*, 34(5):431–436, May 2003.

- [21] J. Diamond. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human pathology*, 35(9):1121–1131, September 2004.
- [22] MD. DiFranco, GO. Hurley, EW. Kay, et al. Automatic gleason scoring of prostatic histopathology slides using multi-channel co-occurrence texture features. In *Proc. of Microscopic Image Analysis with Application in Biology*, 2008.
- [23] S. Doyle, M. Hwang, K. Shah, and A. Madabhushi. Automated grading of prostate cancer using architectural and textural image features. *IEEE International Symposium On Biomedical Imaging*, pages 1284–1287, 2007.
- [24] GS. Dulai, S. Guha, KL. Kahn, et al. Preoperative prevalence of barrett’s esophagus in esophageal carcinoma: A systematic review. *Gastroenterology*, pages 26–33, 2002.
- [25] L. Fei-fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of IEEE computer vision and pattern recognition*, pages 524–531, 2005.
- [26] JC. Felipe, AJM. Traina, and C.T Jr. Retrieval by content of medical images using texture for tissue identification. In *Proc. of the IEEE Symposium on Computer-Based Medical Systems*, pages 175–180, June 2003.
- [27] JF. Flejou. Barrett’s oesophagus: from metaplasia to dysplasia and cancer. *Int. Journal of Gastroenterology and Hepatology*, 54(1):6–12, 2005.
- [28] R. Flipovych and C. Davatzikos. Semi-supervised pattern classification of medical images: application to mild cognitive impairment(mci). *Neuroimage*, 55(3):1109–1119, 2011.
- [29] I. Fogel and G. Sagi. Gabor filter as texture discriminator. *Biological Cybernetics*, 61:103–113, 1989.
- [30] F. Galasso and J. Lasenby. Fourier analysis and gabor filtering for texture analysis and local reconstruction of general shapes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2342–2349, 2009.
- [31] M. Gangeh, A. Ghodsi, and M. Kamel. Dictionary learning in texture classification. In *Proc. of Int. Conf. on Image analysis and recognition*, pages 335–343, 2011.



- [32] JM. Geusebroek, AWM. Smeulders, F. Cornelissen, and H. Geerts. Segmentation of tissue architecture by distance graph matching. *Journal of Cytometry*, 35, 1999.
- [33] Z. Ghahramani. *Advance lecture on machine learning*, chapter Unsupervised learning. 3176. Springer Verlag, 2004.
- [34] GG. Ginsberg. Endoscopic approaches to barrett's oesophagus with high-grade dsplasia/early mucosa cancer. *Best Practice and Research Clinical Gastroenterology*, 22(4):751–772, 2008.
- [35] K. Glatz-Krieger, U. Spornits, A. Spatz, et al. Factors to keep in mind when introducing virtual microscopy. *The European Journal of Pathology*, 448(3):248–255, 2006.
- [36] M. Guillaud, P. Payne, C. Dawe, et al. Quantitative architecture analysis of bronchiol intraepithelial neoplasma. In *Proc. of SPIE BIOS 2000*, January 2000.
- [37] MN. Gurcan, LE. Boucheron, A. Can, et al. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.
- [38] RC. Haggitt. Barrett's esophagus, dysplasia and adenocarcinoma. *Human Pathology*, 25(10):982–993, October 1994.
- [39] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.
- [40] PW. Hamilton, PH. Bartels, D. Thompson, et al. Automated location of dysplastic fields in colorectal histology using image texture analysis. *Journal of Pathology*, 182, 1997.
- [41] RM. Haralick, K. Shanmugam, and I. Disnden. Textural features for image classification. *IEEE Transaction on Systems, Man and Cybernatics*, SMC-3(6):610–621, November 1973.
- [42] H. Helin, M. Lundin, J. Lundin, et al. Web-based virtual microscopy in teaching and standardising gleason grading. *Human Pathology*, 36:381–386, 2005.
- [43] M. Hu, X. Ping, and Y. Ding. Automated cell nucleus segmentation using improved snake. *Int. Conference on Image Processing*, pages 2737–2740, 2004.

- [44] BSG Guidelines in Gastroenterology. Guidelines for the diagnosis and management of barrett's columnar-lined oesophagus. *BSG Guidelines in Gastroenterology*, (28), 2005.
- [45] Aperio Technologies Inc. Aperio imageserver programmer's reference. Technical Report MAN-0070, Revision B, Aperio Technologies Inc., Aperio, 3, The Sanctuary, Eden Office Park, Ham Green, Bristol BS20 0DD, December 2007.
- [46] H. Irshad, S. Jalili, L. Roux, D. Racocaenu, and LJ. Hwee. Automated mitosis detection using texture, sift fetures and hmax biologically inspired approach. In *Proc. on Histopathology Image Analysis in MICCAI'12*, 2012.
- [47] Kourosh. Jafari-Zadeh and Hamid Soltani-Zadeh. Multiwavelet grading of pathological images of prostate. *Biomedical Engieering*, 50(6):697–704, 2003.
- [48] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 2002.
- [49] M. Jondet, R. Agoli-Agbo, and L. Dehennin. Automatic measurementof epithelium differentiation and classification of cervical intraneoplasia by computerized image analysis. *Diagnostic Pathology*, 5(7), 2010.
- [50] A. Karahaliou, S. Skiadopoulos, I. Boniatis, P. Sakellaropoulos, E. Likaki, G. Panayiotakis, and L. Costaridou. Texture analysis of tissue surrounding microcalcifications on mammograms for breast cancer diagnosis. *The British J. of Radiology*, 80:647–656, 2007.
- [51] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. J. of Computer Vision*, 1(4):321–331, 1988.
- [52] K. Kayser. Quantification of virtual slides:approaches to analysis of content-based image information. *Journal of Pathology Informatics*, 2(2):1–12, January 2011.
- [53] K. Kayser, J. Gortler, K. Metse, et al. How to measure image quality in tissue-based diagnosis (diagnostic surgical pathology). *Diagnostic Pathology*, 3(Suppl 1):1–7, July 2008.
- [54] K. Kayser, SA. Hoshang, K. Metze, et al. Texture and object-related automated information analysis in histological still images of various organs. *Journal of Analytical and Quantitative Cytology and Histology*, 30(6):323–335, December 2008.

- [55] SJ. Keenan, J. Diamond, WG. McCluggage, et al. An automatic machine vision for histological grading of cervical intraepithelial neoplasia. *Journal of Pathology*, 192(3):351–362, 2000.
- [56] M. Kerkhof, HV. Dekken, EW. Steyerberg, et al. Grading of dysplasia in barrett’s oesophagus: Substantial interobserver variation between general and gastrointestinal pathologists. *Histopathology*, 50(7):368–378, 2007.
- [57] P. Khurd, C. Bahlmann, P. Maday, et al. Computer-aided gleason grading of prostate cancer histopathological images using texton forests. In *Proc. of IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 636–639, April 2010.
- [58] A. Mohd Khuzi, R. Besar, WMD. Wan Zaki, et al. Identification of masses in digital mammogram using gray level co-occurrence matrices. *Biomedical Imaging and Intervention Journal*, 5(3), 2009.
- [59] P. Kruzinga, N. Petkov, and SE. Grigorescu. Comparison of texture features based on gabor filters. In *Proc. of the 10th Int. Conf. on Image Analysis and Processing*, pages 142–147, September 1999.
- [60] G. Landini and IE. Othman. Architectural analysis of oral cancer, dysplastic and normal epithelia. *Cytometry, Part A*(61A):45–55, 2004.
- [61] S. Lankton. Sparce field method-technical report. Technical report, Georgia Institute of Technology, July 2009.
- [62] AS. Leong and SJ. Leong. Strategies for laboratory cost containment and for pathologist shortage: Centralised pathology laboratories with microwave-simulated histoprocessing and telepathology. *Pathology*, 37:5–9, 2005.
- [63] C. Lepage, B. Rachet, V. Jooste, et al. Continuing rapid increase in esophageal adenocarcinoma in england and wales. *The American Journal of Gastroenterology*, 103:2694–2699, 2008.
- [64] R.A. Lerski, K. Straughan, L.R. Schad, D. Boyle, S. Blumi, and I. Zuna. Mr image texture analysis- an approach to tissue characterization. *Magnetic Resonance Imaging*, 11:873–887, 1993.
- [65] S. Li, L Fang, and S. Member. An efficient dictionary learning algorithm and its application to 3d medical image denoising. *IEEE transaction on biomedical engineering*, 59(2):417–427, 2012.

- [66] Z. Li-jia, Z. Shao-min, and Z. Da-zhe. Medical image retrieval using sift feature. In *Int. Congress on Image and Signal processing*, pages 1–4, 2009.
- [67] CG. Loukas, GD. Wilson, B. Vajnovic, and A. Linney. An image analysis-based approach for automated counting of cancer cell nuclei in tissue sections. *Cytometry Part A*, 55A:30–42, 2003.
- [68] D. Lowe. Object recognition from local scale invariant features. In *Proc. of the 7th Int Conf. on Computer Vision*, pages 1150–1157, 1999.
- [69] D.G Lowe. Distinctive image features from scale-invariant key points. *Int. J. of Computer Vision*, 60(2):91–110, 2004.
- [70] M. Lundin, J. Szymas, E. Linder, et al. A european network for virtual microscopy - design, implementation and evaluation of performance. *The European Journal of Pathology*, 454(4):421–429, 2009.
- [71] DR. Magee, D. Treanor, D. Crellin, et al. Colour normalisation in digital histopathology images. In *Proc. Optical Tissue Image Analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, pages 100–111, 2009.
- [72] J. Malik, S. Belongie, T. Leung, et al. Colour and texture analysis for image segmentation. *J. of Computer Vision*, 43(1):7–27, 2001.
- [73] D. Martens, B. Baesens, J. Vanthienen, and TV. Gestel. Comprehensible credit scoring models using rule extraction from support vector machines. Technical report, University Of Southampton, 2007.
- [74] LO. Martins, AM. Santos, AC. Silva, et al. Classification of normal, benign and malignant tissues using co-occurrence matrix and bayesian neural network in mammographic images. In *Proc. of the Ninth Brazilian Symposium on Neural Networks, SBRN '06*, pages 5–, Washington, DC, USA, 2006. IEEE Computer Society.
- [75] DM. Maru. Barrett’s esophagus: Diagnostic challenges and recent developments. *Annals of Diagnostic Pathology*, 13(2009), 2009.
- [76] K. Masood and N. Rajpoot. Texture based classification of hyperspectral colon biopsy samples using clbp. In *Proc. of the Sixth IEEE Int. conference on Symposium on Biomedical Imaging: From Nano to Macro, ISBI'09*, pages 1011–1014, Piscataway, NJ, USA, 2009. IEEE Press.

- [77] K. Masood, N. Rajpoot, H. Qureshi, et al. Co-occurrence and morphological analysis for colon tissue biopsy classification. In *Proc. of the 4th International Workshop on Frontiers of Information Technology*, Islamabad, Pakistan, December 2006.
- [78] K. Massod and N. Rajpoot. Classification of hyperspectral colon biopsy images: does 2d spatial analysis suffice. In *Ann. British Machine Vision Association*, pages 1–15, 2008.
- [79] T. Matsuyama, K. Saburi, and M. Nagao. A structural analyser for regularly arranged textures. *Computer Graphics and Image Processing*, 18(3):259–278, 1982.
- [80] J. Meyer and Z. Zhang. Wavelet-based image segmentation: Feature recognition in noise-distorted biomedical images. In *Proc. of UC System Wide Bioengineering Symposium*, June 2004.
- [81] S. Mika, C. Schäfer, P. Laskov, D. Tax, and K.R. Müller. *Handbook of Computational Statistics.*, chapter Support Vector Machines. 2nd ed. Springer, June 2012.
- [82] E. Montgomery. Is there a way for pathologists to decrease interobserver variability in the diagnosis of dysplasia? *Arc. Pathol. Lab. Med*, 129:174–176, 2005.
- [83] E. Montgomery, MP. Bronner, and JR. Goldblum. Reproducibility of the diagnosis of dysplasia in barrett’s esophagus: a reaffirmation. *Hum. Pathology*, 32:368–378, 2001.
- [84] A. Montillo, D. Metaxas, and L. Axel. Extracting tissue deformation using gabor filter banks. In A. Amir, A. Amini, and A. Manduca, editors, *Proc. of SPIE*, pages 1–9, 2004.
- [85] S. Naik, S. Doyle, S. Agner, and S. Madabhushi. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. *Int. Symposium on Biomedical Imaging*, pages 284–287, 2008.
- [86] S. Naik, S. Doyle, M. Feldman, et al. Gland segmentation and computerized gleason grading of prostate histology by integrating low, high-level and domain specific information. *Conference of Microscopic Image Analysis with Application in Biology*, 2008.
- [87] RD. Odze. Diagnosis and grading dysplasia in barrett’s oesophagus. *Journal of Clinical Pathology*, pages 1029–1038, 2006.

- [88] American Board of Pathology. Instructions and information for candidate for certifying examination. online <http://www.abpath.org>, PO Box 25915, Tampa, Florida, May 2012.
- [89] T. Ojalo, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.
- [90] RC. Orlando. *Reflux Esophagitis*, pages 1235–1263. Lippicott Williams and Wilkins, 3rd edition, 1999.
- [91] AH. Ormsby, RE. Petras, WH. Henricks, et al. Observation variation in the diagnosis of superficial oesophagus adenocarcinoma. *Int. Journal of Gastroenterology and Hepatology*, 2002.
- [92] C. Palm. Color texture classification by integrative co-occurrence matrices. *Pattern Recognition*, 37(5):962–976, 2004.
- [93] V. Punys, J. Puniene, R. Jurkevicius, et al. Myocardium tissue analysis based on textures in ultrasound images. In R. Engelbrecht, A. Geissbuhler, C. Lovis, and G. Mihalas, editors, *Connecting Medical Informatics and Bio-informatics: Proc. of MIE2005*, pages 435–440. European Federation for Medical Informatics, IOI Press, 2005.
- [94] JR. Quinlan and RL Rivest. Inferring decision tree using the minimum description length principle. *Information and computation*, 80:227–248, 1989.
- [95] H. Qureshi. *Meningioma Classification using an adaptive discriminant wavelet packet transform*. PhD thesis, Computer Science, University of Warwick, October 2009.
- [96] H. Qureshi, O. Sertel, N. Rajpoot, et al. Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. In *Proc. of Medical Image Computing and Computer-assisted Intervention*, pages 196–204, New York, USA, September 2008. Springer Berlin/Heidelberg.
- [97] BJ. Reid, RC. Haggitt, CE. Rubin, et al. Observer variation in the diagnosis of dysplasia in barrett’s esophagus. *Human Pathology*, 19(2):166–178, February 1988.
- [98] E. Reinhard, M. Ashikhmin, B. Gooch, et al. Color transfer between images. *IEEE Computer Graphics and Applications*, pages 34–41, October 2001.

- [99] TW. Rice, JE. Mendelin, and JR. Goldblum. Barrett's oesophagus: Pathologic considerations and implications for treatment. *Thoracic and cardiovascular surgery*, 17(4):292–300, 2005.
- [100] Haralick. RM. Statistical and structural approach to texture. *Proc. of IEEE*, 67(5):786–804, May 1979.
- [101] R. Rodrguez, TE. Alarcon, and O. Pacheco. A new strategy to obtain robust markers for blood vessels segmentation by using the watersheds method. *Computers In Biology And Medicine*, 35:665–686, 2005.
- [102] MA. Roula, A. Bouridane, and F. Kurugollu. An evolutionary snake algorithm for segmentation of nuclei in histopathological images. *Int. Conference on Image Processing*, pages 127–130, 2004.
- [103] AC. Ruifrok and DA. Johnston. Quantification of histological staining by color deconvolution. *Analytical Quantitative Cytology and Histology*, 23:291–299, 2001.
- [104] LA. Ruiz, A. Fdez-sarra, and JA. Recio. Texture feature extraction for classification of remote sensing data using wavelet decomposition: a comparative study. In *International Archives of Photogrammetry and Remote Sensing.*, pages 1682–1750, 2004.
- [105] O. Sertel, J Kong, G. Lozanski, et al. Texture classification using nonlinear color quantization: Application to histopathological image analysis. In *Proc. of IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 597–600. IEEEExpore, April 2008.
- [106] NJ. Shaheen and JE. Richter. Barrett's oesophagus. *The Lancet*, 373:850–861, March 2009.
- [107] JK. Shuttleworth, AG Todman, and MK. Bennet. Colour texture using co-occurrence matrices for classification of colon cancer. In *Proc. of IEEE Canadian Conference On Electrical And Computer Engineering*, pages 1134–1139, 2002.
- [108] P. Singh and P. Compton. Evolution oriented semi-supervised approach for segmentation of medical images. In *Proc. of Int. Cont. of intelligent sensing and information processing*, pages 77–81, 2004.
- [109] B. Snape. Grading dysplasia in barrett's oesophagus using computer vision. Master's thesis, School of Computing, University of Leeds, 2008/2009.

- [110] M. Solaymani-Dodaran, RFA. Logan, J. West, et al. Risk of oesophageal cancer in barrett's oesophagus and gastro-oesophagus reflux. *Int. Journal of Gastroenterology and Hepatology*, 53, 2004.
- [111] SJ. Spechler. Barrett's esophagus. *The New England Journal of Medicine*, 346(11), 2002.
- [112] J. Sudbo, A. Bankfalvi, M. Bryne, et al. Prognostic value of graph theory-based tissue architecture analysis in carcinomas on tongue. *Laboratory Investigation*, 80(12):1881–1889, 2000.
- [113] J. Sudbo, R. Marcelpoil, and A. Reith. New algorithms based on voronoi diagram applied in a pilot study on normal mucosa and carcinomas. *Analytical Cellular Pathology*, 21:71–86, 2000.
- [114] R. Susomboon, D. Raicu, J. Furst, et al. A co-occurrence texture semi-invariance to direction, distance and patient size. In *Proc. of SPIE Medical Imaging Conference*, 2008.
- [115] A. Tabesh, M. Teverovskiy, H-Y. Pang, et al. Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE Transactions on Medical Imaging*, 26(10):1366–1378, 2007.
- [116] G. Tamura, S. Mikawa, and N. Monma. Secretory carcinoma coexistent with mucinous carcinoma in the breast: Report of a case. *Acta Pathology Japan*, 39:593–598, 1989.
- [117] C. Tang and Y. Dong. Automatic registration based on improved sift for medical microscopic sequence images. In *Int. Symposium on Intelligent information technology application*, pages 580–583, 2008.
- [118] DL. Thiele, C. Kimme-Smith, TD. Johnson, McCombs M., and LW. Bassat. Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes. *Med Phys*, 23(4):549–555, Apr 1996.
- [119] D. Treanor, CH. Lim, DR. Magee, and AJ. Bulpitt. Tracking with virtual slides: A tool to study diagnostic error in histopathology. *Histopathology*, 55(1):37–45, 2009.
- [120] AJ. Viera and JM. Garret. Understanding interobserver agreement: The kappa statistics. *Family Medicine*, 37(5):360–363, May 2005.



- [121] KK. Wang and Sampliner RE. Updated guidelines 2008 for the diagnosis, surveillance and therapy of barrett's esophagus. *American Journal of Gastroenterology*, 103, 2008.
- [122] Y. Wang, D. Crookes, Ji. Diamond, et al. Segmentation of squamous epithelium from ultra-large cervical histological virtual slide. In *Proc. of the 29th Annual Int. Conference of IEEE Engineering in Machine and Biology Society*, pages 775–778, 2007. August.
- [123] Y. Wang, K. Itoh, N. Taniguchi, et al. Studies on tissue characterization by texture analysis with co-occurrence matrix method using ultrasonography and ct imaging. *Journal of Medical Ultrasonic*, 29(4):211–223, 2002.
- [124] B. Weyn, GVE. Wowver, AV. Daele, et al. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry*, 33:32–40, 1998.
- [125] PF. Whelan and D. Molley. *Machine vision algorithms in java: ttechniques and implementation*, chapter Texture analsis. Springer-Verlag London, 2002.
- [126] JN. Wolfe. Wolfe mammographic parenchymal patterns and breast cancer risk. *American Journal of Radiology*, 126:1130–1139, 1976.
- [127] J. Wu and X. Zhang. A pca classifier and its application in vehicle detection. In *Proc. of Int. Joint Conference on Neural Networks*, pages 600–604, Washington, DC, July 2001. IEEE.
- [128] Jun Xu, Rachel Sparks, Andrew Janowczyk, John E. Tomaszewski, Michael D. Feldman, and Anant Madabhushi. High-throughput prostate cancer gland detection, segmentation, and classification from digitized needle core biopsies. In *Proc. of the 2010 international conference on Prostate cancer imaging: computer-aided diagnosis, prognosis, and intervention*, MICCAI'10, pages 77–88, Berlin, Heidelberg, 2010. Springer-Verlag.
- [129] X. Yang, H. Li, and X. Zhou. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift and kalman filter in time lapse microscopy. *IEEE Transaction on Circuits and Systems*, 2006.
- [130] W. Yi-Ying, C. Shao-Chien, et al. Color-based approach for automated segmentation in tumor tissue classification. *Engineering in Medicine and Biology Society*, 2007.

- 
- [131] F. Yousef, C Cardwell, MM. Cantwell, et al. The incidence of esophageal cancer and high-grade dysplasia in barrett's esophagus: A systematic review and meta-analysis. *American Journal of Epidemiology*, 168(3):237–249, 2008.
- [132] F. Yu and H. Ip. Semantic content analysis and annotation of histological images. *Computers in Biology and Medicine*, 38(6):635–494, June 2008.
- [133] Shaotong Zhang, Yiqian Zhan, and Dimitris N. Metaxas. Deformable segmentation via sparse representation and dictionary learning. *Medical image analysis*, 16(7):1385–1396, October 2012.
- [134] X. Zhu and Andrew B. Goldberg. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, chapter Introduction to Semi-Supervised Learning. 6. Morgan & Claypool, 2009.